Evaluation of Four Reports on Contamination of the Athabasca River System by Oil Sands Operations

Prepared By:

Water Monitoring Data Review Committee

Peter Dillon, George Dixon, Charles Driscoll, John Giesy, Stuart Hurlbert, Jerome Nriagu

Prepared For:

Government of Alberta

March 7, 2011

EXECUTIVE SUMMARY

In 2008, a team of scientists, led by Drs. David Schindler and Erin Kelly of the University of Alberta [called Kelly et al. throughout the report; *et al.* is Latin for "and others"], conducted an intensive short-term study to determine if oil sands operations in the Athabasca region of Alberta were releasing pollutants to the watershed of the Athabasca River and the river itself. The findings of the study were published in two articles in a prestigious scientific journal. These articles stated that the oil sands development is a greater source of contaminants than was previously recognized. Many of the pollutants that they found in snow pack and river samples are toxic at low concentrations (trace metals), while others could harm fish embryos in the rivers (polycyclic aromatic compounds, PAC). They also indicated that current monitoring programs throughout the oil sands region are inadequate to determine impacts of these chemicals.

Since their publication in 2009 and 2010, these articles have received considerable attention and generated controversy. In September 2010, Premier Ed Stelmach of Alberta announced that he would work with Prof. Schindler to set up an independent panel of experts in the field of water pollution and its effects on aquatic systems. The panel, called the *Water Monitoring Data Review Committee*, was instructed to review the articles by Kelly et al. and reports by Alberta Environment and the Regional Aquatics Monitoring Program (RAMP). They were charged with examining study designs, data, and statistical approaches, to determine if the conclusions among these reports were consistent and comparable.

Although many reports, studies and articles have been written on environmental impacts of the oil sands, the Committee decided to focus on four documents. These were the two articles by Kelly et al., one Alberta Environment report by T. Hebben and one RAMP Review report.

The focused, short term sampling campaign used by Kelly et al. was adequate for estimating short-term inputs to the watershed in the region of the oil sands development and potential impacts to the aquatic ecosystem. The Alberta Environment study included monitoring at a limited number of stations, and was not specifically intended to determine impacts from the oil sands operations. The RAMP program has many monitoring sites, but the low sampling frequency each year limits this program's ability to determine impacts from oil sands operations.

Monitoring of chemicals in river systems, particularly large complex systems such as the Athabasca River and its tributaries, is not easy. Concentrations can vary substantially over time (season, flow rate) and over space within the system. As a result, it is difficult to use a few measurements of concentrations of contaminants in river water and fine surface sediments to assess environmental impacts. The water and fine surface sediments in a stream or river represent materials in transit and will always contain, at any given location, materials more or less recently arrived from upriver. Even with constant inputs from oil sands operations, one would not necessarily expect an increase over time in concentrations of contaminants in water and sediment in rivers, even at points downstream of these operations.

Conclusions in Kelly et al.'s articles and the report by Alberta Environment are based on monitoring chemicals that might have an effect, as opposed to monitoring actual effects on organisms (biological monitoring). They mostly assess risks by comparing chemical concentrations in their samples to either water quality guidelines (for trace metals) or published concentrations that may cause an effect on certain aquatic organisms (for polycyclic aromatic hydrocarbons, PAHs). The need for data that can be used in comprehensive risk assessment cannot be over-emphasized.

Kelly et al.'s study has limitations on its ability to estimate the amount of contaminants entering the river and its watershed from the oil sands activities. There were large disparities between the estimated emissions from oil sands operations and Kelly et al.'s estimated deposition near these facilities. Their results, therefore, carry the implication that considerably more particulate matter and trace metals are being released from the oil sands facilities than is being reported in the National Pollutant Release Inventory, or that there are other airborne sources. Recent studies show that levels of polycyclic aromatic hydrocarbons (PAHs) in sediments of the Athabasca delta and mercury in the eggs of birds nesting there have been increasing, as have arsenic concentrations in the sediments of Lake Athabasca. The articles by Kelly et al. have served to point out some of the gaps in data and understanding in linking sources and effects. In particular, at the time they were done, there were no other studies on atmospheric deposition of contaminants in the oil sands area, especially to terrestrial ecosystems.

The Alberta Environment report was intended to document long-term trends in about 100 chemicals in the Athabasca River, over time and along the river. It was not intended to assess impacts of the oil sands on the river. The limitations in quantity and accuracy of long-term trace metal data made it difficult for the author to apply trend tests to them. The report did state, however, that concentrations of a few trace metals in the river exceeded water quality guidelines on some occasions.

Although Alberta Environment monitored polycyclic aromatic hydrocarbons (PAH) in the river, they were mostly undetected in the river samples. Therefore, the author did not conduct trend analyses on PAH. A fair amount of effort and resources have apparently been devoted to generate the mostly non-detect data for PAH and other organic contaminants. Failure to detect PAHs in the sample cannot be a reflection of available analytical technology between 1990 and 2007. Although it is likely that concentrations in the river are generally low, it appears the laboratories Alberta Environment used do not have the capability of measuring the low concentrations of PAC found in the water. The Committee believes there were deficiencies in the sampling design and methodology for this study. The development of appropriate detection limits based on the potential for toxic effects would have made the study more relevant. Some aquatic organisms can accumulate PAHs to concentrations that would result in harm even from very small concentrations in the environment.

The Regional Aquatics Monitoring Program (RAMP) has a very extensive monitoring design. But the Committee believes the program is spending large amounts of time and resources on obtaining water quality data that are difficult to interpret because the systems they are monitoring are large, complex and variable, and their sampling frequencies are too low and the sampling locations are not adequate to account for this. Although many different trace metals are measured in the samples collected, only a few of these fit the program's monitoring criteria and hence the others are not reported. Data for all trace metals measured in water samples should be included in future reports by RAMP.

To assess contaminants in the rivers and watershed that are contributed by oil sands operations is made difficult by the scarcity of information on natural historical background levels and on true reference sites. This is generally understood, but must be kept in mind where the various reports refer to "upstream" versus "downstream," "background" versus "near development," and "test" versus "baseline" sites or conditions. There are at least four underlying causes for absence of adequate reference sites. First, no monitoring programs were put in place until well after oil sands operations began. Second, once they were in place occasional changes in analytical labs used and in detection limits have complicated assessment of temporal trends. Third, locations and amounts of natural inputs, especially by groundwater inflows, are poorly known. In many cases such natural inputs are upriver of oil sands operations. And fourth, aerial dispersal and subsequent deposition of contaminants generated by oil sands operations will often be southward, i.e. upriver of Fort McMurray, as winds are frequently from the north.

There are several reasons for the apparent differences of opinion about whether oil sands contaminants are derived from natural sources or the oil sands industry ("the controversy"). Two main reasons are: 1) each of the databases in these reports is limited in terms of quality, quantity and/or lack of spatial and temporal resolutions. They cannot scientifically justify all of the inferences that have been reported; and 2) each study used different reference sites to compare levels of contaminants attributed to natural sources and human activities.

Taking into consideration all data and critiques, we generally agree with the conclusion of Kelly et al. that PACs and trace metals are being introduced into the environment by oil sands operations. However, their estimates of PAC deposition rates must be regarded as only approximate and preliminary in nature. We agree with Kelly et al. that it is improbable that the snowpack-deposited contaminants could have resulted from wind erosion of bitumen outcrops or bitumen-containing soils in undisturbed landscapes – especially under snow-cover. Information on PAC concentrations in water provided by Kelly et al. is less conclusive. While many of the differences they document are consistent with large inputs of contaminants from oil sands operations, their water data do not allow for a quantitative analysis of the relative contributions of natural loadings and those due to oil sands operations. The comments above are equally valid for trace metals.

We generally see no conflict among the conclusions from Kelly et al., the Alberta Environment report, and the RAMP reports. Although the Alberta Environment trace metals data from before 2004 are invalid, most of the data in these reports are valid. The Kelly et al. study was highly focused, the Alberta Environment study examined long-term trends, and RAMP elucidated patterns in water quality and inputs over space and time, but at a low sampling frequency. Because all of these studies had a different focus, their conclusions generally did not conflict, despite occasional differences in interpretations or neglected patterns.

We think Kelly et al.'s study, in spite of some uncertain statements on loadings and risks, has been important in pointing out deficiencies in current monitoring programs in the oil sands area. We believe it is in the best interests of the public and the oil sands industry to make sure all monitoring programs are conducted with scientific rigor and oversight.

The studies by Kelly et al. have served to focus attention on some critical issues that can be resolved in a new monitoring program now being designed by a committee set up by the Alberta Minister of the Environment. This program can build on elements and concepts of the three existing programs to address the issues of whether the releases from oil sands production are causing adverse effects on aquatic and terrestrial organisms. This monitoring should consider effects in tributaries, especially during critical periods of flow in the river. The accumulation of residues in the delta, Lake Athabasca and their biota also merit special attention, with expanded biological monitoring and focused scientific investigations to assess risk.

TABLE OF CONTENTS

LIST OF TABLES AND FIGURES

ACRONYMS AND ABBREVIATIONS

AENV	Alberta Environment
AM	Arithmetic Mean
ANOVAs	Analysis of Variance
CALA	Canadian Association for Laboratory Accreditation
CCME	Canadian Council of Ministers of the Environment
CEMA	Cumulative Environmental Management Association
CEQG	Canadian Environmental Quality Guidelines
CONRAD	Canadian Oilsands Network for Research and Development
CVAFS	Cold Vapor Atomic Fluorescence Spectrometry
DL	Detection Limit
EC	Environment Canada
ERCB	Energy Resources Conservation Board
GM	Geometric Mean
ICP-AES	Inductively Coupled Plasma Atomic Emission Spectroscopy
ICP-MS	Inductively Coupled Plasma Mass Spectrometry
ISO/IEC	International Organization for Standardization/International Electrotechnical Commission
LTRN	Long-Term River Network
MDN	Mercury Deposition Network
ng/L	Nanogram per Litre
NPRI	National Pollutant Release Inventory
P Value	Probability
PAC	Polycyclic Aromatic Compound
PAH	Polycyclic Aromatic Hydrocarbon
PM2.5	Fine Particulate Matter 2.5 Microns in Diameter
PMD	Polyethylene Membrane Device
QA/QC	Quality Assurance/Quality Control
RAMP	Regional Aquatics Monitoring Program
RAMP IT	Regional Aquatics Monitoring Program Implementation Team
RSC	Royal Society of Canada
SCO	Synthetic Crude Oil
SE	Standard Error
SOP	Standard Operating Procedure
$\mu g/m^2$	Microgram per Square Metre
USGS	United States Geological Society
VOC	Volatile Organic Carbon
WBEA	Wood Buffalo Environmental Association
WMDRC	Water Monitoring Data Review Committee
WQI	Water Quality Index

1.0 INTRODUCTION

A group of scientists led by Dr. Erin Kelly *et al.* and Prof. David Schindler of the University of Alberta published two articles in a scientific journal[1],[2]. These articles suggested that the oil sands industry is releasing trace metals and polycyclic aromatic compounds (PACs) that are potentially harmful to the Athabasca River and its watershed. Thirteen trace metals were reported on in one of these articles. These are called "priority pollutants" by the U.S. Environmental Protection Agency. These metals can be toxic and last a long time in the environment.

Polycyclic aromatic compounds could harm fish populations and other aquatic organisms, and possibly people using the river as a water supply. Kelly *et al.* measured all of these contaminants in snowpack to estimate their deposition from the atmosphere from oil sands activities and local emissions. They also measured these contaminants in the water of the Athabasca River and its tributaries in the oil sands area, and in Lake Athabasca and the Athabasca Delta.

Alberta government reports suggest that the majority of oil sands-related pollutants detected downstream of oil sands activities are derived from the release from natural bitumen outcrops (hydrocarbons) resulting from local geology. Work to assess the relative importance of various factors on water quality in the Athabasca River system is complex and ongoing. Kelly *et al.*, however, have doubts about whether Alberta's monitoring programs are adequate to determine the amounts and sources of these chemicals that may be entering the watershed and river. They also suggested that the monitoring programs of RAMP and Alberta Environment are unable to assess the impacts of oil sands activities on the water, aquatic life, and other natural resources very well.

Premier Ed Stelmach of Alberta announced on September 24, 2010 that the government would work with Prof. Schindler to put together an independent committee of six scientists who are experts in the field of water pollution and its effect on aquatic ecosystems. The committee, called the *Water Monitoring Data Review Committee*, was given a mandate to review data and conclusions from the work of Kelly *et al.* and the Alberta monitoring programs. Prof. Schindler suggested three of the committee members and the Government of Alberta suggested the other three members. The Minister of Alberta Environment announced the committee members on October 7, 2010.

---

[1] Kelly, E.N., J.W. Short, D.W. Schindler, P.V.Hodson, M. Ma, A.K. Kwan and B.L. Fortin. Oil sands development contributes polycyclic aromatic compounds to the Athabasca River and its tributaries. Proc. Nat. Acad. Sciences, 106:52, p 22346-22351.

[2] Kelly, E.N., D.W. Schindler, P.V. Hodson, J.W.Short, R. Radmanovich, and C.C. Nielsen. Oil sands development contributes elements toxic at low concentrations to the Athabasca River and its tributaries. Proc. Nat. Acad. Sciences, 107:37, p 16178-16183.

1.1 Members of the Water Monitoring Data Review Committee

The names and brief biographies of the six members of the committee follow in alphabetical order:

- *Peter Dillon:* Dr. Dillon is the director of the Water Quality Centre at Trent University and a professor in the Environmental and Resource Studies and Chemistry departments. His expertise is in the field of watershed biogeochemistry.

- *George Dixon:* Dr. Dixon is the vice-president of research and professor of biology at the University of Waterloo. His specialty is effects of toxic chemicals, including metals and oil sands process water, on aquatic organisms, principally fish.

- *Charles Driscoll:* Dr. Driscoll's research focuses on environmental chemistry, biogeochemistry (the study of cycles of chemical elements and their interactions with living things) and water quality response to ecosystem disturbance. He is a university professor of environmental systems engineering at Syracuse University in New York State and is a member of the U.S. National Academy of Engineering.

- *John Giesy:* Dr. Giesy is a professor and Canada research chair in Environmental Toxicology in the Department of Veterinary Biomedical Sciences and Toxicology Centre at the University of Saskatchewan. He is regarded as one of the world's eminent ecotoxicologists (study of toxic effects, caused by natural or synthetic pollutants, to entire ecosystems) and a fellow of the Royal Society of Canada.

- *Stuart Hurlbert:* Dr. Hurlbert is the former director of the Center for Inland Waters and professor emeritus at San Diego State University. He specializes in human population issues, biostatistics and limnology.

- *Jerome Nriagu:* Dr. Nriagu is a professor in the School of Public Health as well as a research professor in the Center for Human Growth & Development at the University of Michigan. A fellow of the Royal Society of Canada, Dr. Nriagu is an expert who has published extensively on toxic trace metals in the Canadian environment.

1.2 Purpose of the Water Monitoring Data Review Committee (WMDRC)

The charges to the WMDRC were to:

- Review the two technical articles published by Dr. Kelly *et al.* in the Proceedings of the National Academy of Sciences USA, and 2) surface water data and reports from RAMP and Alberta Environment, to compare findings and determine if the quality of data is adequate to support conclusions from each of the two groups.
- Examine study designs, methodology, data handling, and statistical techniques for the studies of both the Kelly *et al.* and the Government of Alberta.
- Determine if the conclusions from the University of Alberta scientists and the Government of Alberta are consistent and comparable.

- If there are differences in the data supporting the conclusions, are they important? Are there any gaps in the data, and are these relevant?
- Identify the risks and impacts in the use of the data.
- Other reports and articles can be reviewed to find more information to help with this report.

1.3 The Approach of the Water Data Monitoring Review Committee

Our approach was to have teleconference meetings regularly, instead of in-person, and have representatives from Prof. Schindler's team, Regional Aquatics Monitoring Program, Alberta Environment, Wood Buffalo Environmental Association, industry and Environment Canada, answer specific questions about their work, both written and verbally (see Table 1). First Nations representatives were also invited to tell us about their impressions of oil sands impacts. The WDMRC also toured the oil sands area to get a better understanding of the issues and listen to staff from the mining industry, regulators, Alberta Environment, and RAMP, as well as from Dr. Kelly and Prof. Schindler of the University of Alberta. The committee intended to determine whether conclusions from the two groups are warranted based on the data.

We were also provided with a huge volume of material in the form of published and unpublished reports, raw data, and scientific papers to use in our work. Because of the limited amount of time we had to produce the report, the committee decided to focus its analysis on four reports. These were the two papers in the *Proceedings of the National Academy of Sciences* in 2009 and 2010 on deposition rates and concentrations of PACs and trace metals in the Athabasca River region (by Kelly *et al.*); a major report from Alberta Environment (AENV) in 2009 titled *Analysis of Water Quality Conditions and Trends for the Long-Term River Network: Athabasca River, 1960-2007* and another major report in 2010 titled *Regional Aquatics Monitoring Program 2009 Technical Report* (RAMP). These studies had different objectives, are based on markedly different data sets, and asked different questions of those data sets and hence can be considered to be representative of many water quality reports addressing local impacts of the oil sands industry. It should be emphasized that atmospheric inputs were considered in the papers by Kelly *et al.* 2009 and 2010 but were not addressed in the reports by RAMP and AENV.

Table 1 Schedule of meetings and speakers interviewed

DATE	MEETING	SPEAKER
December 6th and 7th, 2010	1) Mining Operations Overview	Suncor
	2) Science & Regulatory Overview	Alberta Environment
	3) RAMP Overview	RAMP and Hatfield Consultants
	4) Kelly/Schindler Overview	Erin Kelly and David Schindler
December 16, 2010	1) Wood Buffalo Environmental Association Overview	WEBEA
January 7, 2011 – Questions and Answer Sessions with:	1) Kelly/Schindler	Erin Kelly and David Schindler
	2) Alberta Environment	
	3) RAMP	Hatfield Consultants and RAMP
January 21, 2011	1) Fort McKay First Nation	
February 3, 2011	1) Environment Canada	

1.4 The Athabasca Oil Sands

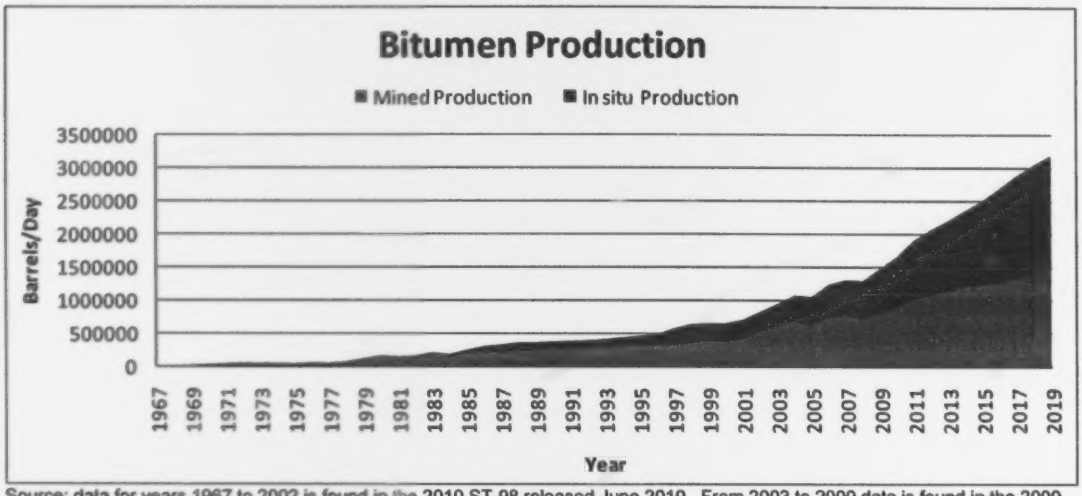
The deposits of heavy oil known as the oil sands lie in northeastern Alberta, and occupy an area of 142,000 km^2 (54,000 mi^2) (ERCB 2010). The area contains the two largest rivers in the province, the Peace and Athabasca rivers. The Athabasca River and some of its tributaries flow through natural oil sands formations downstream of Fort McMurray, the largest settlement in the area.

The oil sands in Alberta are the second largest reservoir of oil in the world (after Saudi Arabia). They are estimated to contain 27 billion m^3 (170 billion barrels) of recoverable oil. About 20% of this oil can be extracted with surface mining. The remainder will require *in-situ* processes, which would result in less land disturbance. At present, there are 22 producing heavy oil operations in the Athabasca oil sands. Thirty-nine projects are either experimental, proposed or under construction. Four of the proposed projects would be surface mining, while the rest would be in-place operations. Many more are either proposed or producing in the Cold Lake and Peace areas of Alberta (Alberta Energy 2010).

The first commercial oil sands operation began in 1967. Capital investment in oil sands projects began to accelerate in the 1990s. Actual investments exceeded $1 billion in the 1990s and reached $4.2 billion in 2000. The rapid growth in investment was sustained during the last decade and averaged over $16 billion per year from 2006 to 2008 (Royal Society 2010). The increase in investment closely matched synthetic crude oil (SCO) production. Approximately 17% of the estimated established bitumen reserves in the Athabasca oil sands region were under active development as of the end of 2008, and 3% of the estimated established bitumen reserves of the Athabasca oil sands region has been extracted by the end of 2008.

With increasing development of the oil sands, there has been a growing public concern about impacts on the environment. There are several potential sources of air pollutants from the operations including surface mining, stationary fuel combustion, flaring, industrial processes, venting, on-site transportation, burning of natural gas to generate steam required for *in-situ* bitumen extraction processes, fugitive dusts tailing piles, and others. Volatile contaminants may also originate from tailing ponds, out-gassing from bitumen at mine faces and as fugitive emissions wherever hydrocarbons are handled. While the point sources of emissions (notably combustion and upgrading sources) may be regulated using the permit process, the non-point sources are notoriously difficult to estimate and control.

Through the years, the public has been concerned about the amount of water being taken out of the Athabasca River for the oil sands industry. Between 2.2 and 5 barrels of river water are withdrawn to produce each barrel of synthetic crude oil. For *in-situ* operations, up to half a barrel of fresh water is required to produce each barrel of bitumen. The amount of water permitted to be withdrawn from the river for all oil sands projects – existing and future – is less than 3% of its average annual flow. During periods of low river flow, Alberta Environment limits water consumption to 1.3% of annual average flow. As a result, industrial users at times will be restricted to less than half of their normal requirement given current approved development and current flows in the river (Alberta Energy 2010).



Source: data for years 1967 to 2002 is found in the 2010 ST-98 released June 2010. From 2003 to 2009 data is found in the 2009 ST-43 released August 2010 and the February 2010 ST-53. From 2010 to 2019 data is found in the 2010 ST 98 released June 2010

Figure 1 Total bitumen production

Water pollution is also a major concern for people in the area. Although tailings ponds are not allowed to discharge into local water bodies, impacts on water quality could result from:

- Leaks and seepage from oil sands operations;
- Spills of mine-impacted waters or chemicals used on site;
- Deposition of air-borne contaminants into aquatic ecosystems via precipitation and runoff;
- Dust and runoff from mining sites;
- Extraction of groundwater and drawdown of aquifers linked to surface water bodies;
- Alteration in the hydrology of aquatic environments; and
- Acidification from air emissions
- Licenced air and water discharges (RAMP website 2010).

1.5 Monitoring Programs to Assess Environmental Impacts

The large quantities of air and water pollutants emitted by individual operations and the large number of these sources are grounds for concern about adverse impacts on local air and water quality. In response to such concerns, Environment Canada established two long-term monitoring sites on the Athabasca River, first at the town of Athabasca in 1960 and then Old Fort in 1978. In 1975, Alberta Environment (AENV) set up the Alberta Oil Sands Environmental Program with the aim of identifying the potential long-term impacts of oil sands development. In 1987, Albert Environment took over the operation of the two Environment Canada stations as part of the provincial Long-term River Network (LTRN) and subsequently added two additional

monitoring sites near Hinton in 1999 and at Fort McMurray in 2002. AENV has recently initiated extensive and integrated monitoring within the Muskeg River watershed, as well as a contaminant loading study throughout the oil sands region and historical sediment quality assessments (coring studies). The rapid expansion of oil sands development in the last 20 years led to the establishment of many research and monitoring organizations and programs to collect data on water and air quality in the oil sands area. These include, among others, the Cumulative Environmental Management Association (CEMA), Wood Buffalo Environmental Association (WBEA), Canadian Oil Sands Network for Research and Development (CONRAD), and the Regional Aquatics Monitoring Program (RAMP).

The Regional Aquatics Monitoring Program (RAMP), an industry-funded multi-stakeholder initiative, was established with a mandate to "determine, evaluate and communicate the state of the aquatic environment and any changes that may result from cumulative resource development within the Regional Municipality of Wood Buffalo" (RAMP 2009). The Cumulative Environmental Management Association (CEMA) was established to develop guidelines and management frameworks on how best to reduce cumulative environmental effects due to industrial development. The Wood Buffalo Environmental Association (WBEA), another industry funded organization, was established to monitor and provide information on air quality and air-related environmental impacts in the Municipality of Wood Buffalo. In addition, individual oil sands companies undertake regular or periodic water quality monitoring in streams and rivers near their operations to satisfy permit requirements. Alberta Sustainable Resource Development monitors and manages the fisheries resource in the region and Water Survey of Canada maintains several hydrology stations in the area. Besides these routine monitoring programs, a number of universities and government laboratories conduct periodic studies on specific aspects of local aquatic ecosystems and their response to oil sands development. With so many organizations and agencies conducting uncoordinated measurements with diverse objectives using a wide variety of methods, a large database on water (and to a limited extent, air, sediments, and aquatic organisms) has been generated. The data are of variable quality, which can lead to contradictory and conflicting conclusions concerning environmental impacts.

In 2008, a team of scientists led by Dr. Erin Kelly and Prof. David Schindler of the University of Alberta conducted studies which led to the publication of two articles in the *Proceedings of the National Academy of Sciences*. These articles suggested that the oil sands industry is releasing trace metals and polycyclic aromatic compounds (PACs) at levels that are potentially harmful to the Athabasca River and its watershed. They measured the contaminants in snowpack to estimate their deposition from the atmosphere from oil sands activities and local emissions. They also measured these contaminants in the water of the Athabasca River and its tributaries in the oil sands area, and in Lake Athabasca and the Athabasca Delta. Their results contradicted statements from the Regional Aquatics Monitoring Program and other government studies to the effect that the majority of oil sands-related pollutants detected downstream of oil sands activities were derived from the weathering of bitumen outcrops along the river and its tributaries. The work by Kelly *et al.* cast some doubt on whether Alberta's monitoring programs are adequate to determine the amounts and sources of these chemicals that may be entering the watershed and river. They also questioned whether the RAMP and Alberta Environment monitoring programs are scientifically rigorous and able to assess adequately the impacts of oil sands activities on the water, aquatic life, and other natural resources.

1.6 Is There a Controversy?

A joint Canada-Alberta-NWT government study for the Northern Rivers Ecosystem Initiative (NREI) (Brua *et al.* 2003) found that the tributaries passing through the Fort McMurray oil sands regions as well as downstream and the Delta of Athabasca River and western Lake Athabasca contained significant levels of PAHs, which they said were likely derived from natural bitumen in the region. They showed that on occasion, some PAHs exceeded the Interim Sediment Quality Guidelines. Other NREI scientists (Parrot *et al.* 2004) also found slight differences in indicators of fish reproductive capacity, such as decreased sex steroid production by testes and ovaries of slimy sculpin) in fish exposed to oil sands along the Steepbank River near anthropogenic disturbance, as well metabolic indicators, such as mixed function monooxygenase enzymes, in livers of slimy sculpin between oil sands and reference sites.

Based on an extensive review of the literature and some analysis of data from other sources, Timony and Lee (2009) found increased levels of PAH when sites downstream of industry were compared with sites upstream of industry. They suggested that the concentrations of PAHs in sediment, mercury in fishes, arsenic in water and sediment and of criteria air contaminants such as fine particulates, volatile organic compounds, and sulphur dioxide had increased significantly over time and that increased rates of fish abnormalities were being observed by local fishermen. They claimed that the levels of PAHs, mercury, and arsenic in the lower Athabasca River system and criteria air pollutants around the oil sands operations had reached levels to be of some concern with regard to human and ecosystem health.

The two articles by Kelly *et al.* (2009, 2010) reported that deposition from the atmosphere and concentrations of polycyclic aromatic compounds (organic chemicals typically found in fossil fuels) and certain trace metals were higher in snowpack and in streams near the oil sands plants than at locations farther away from the oil sands development. They claimed that the inputs of these toxic substances from these activities (via the air and water) are having significant impacts on the Athabasca River, its tributaries and watershed.

In contrast to the studies above, the RAMP's monitoring data were interpreted to suggest that "there were no detectable regional changes in aquatic resources related to oil sands development," although there were a few exceptions at particular locations.

Recently, Alberta Environment concluded that "surface water quality conditions of the lower Athabasca River reflect both natural and anthropogenic influences". This conclusion is based on (a) known and predicted natural and industrial sources of contaminants to the river; (b) results of long-term surface and groundwater monitoring; and (c) preliminary results of new monitoring and research.

An underlying current in the differing interpretations of the trace metals data is the definition of baseline versus background levels. Kelly *et al.* regarded all areas more than 50 km away from the site nearest the upgraders as background for their snow data. They also regarded concentrations in tributaries in areas of the watershed with less than 25% development as being indicative of background levels. It is unlikely that ecosystems with true background concentrations of trace metals exist anywhere near a major industrial development such as the oil

sands facilities, which have been in existence for about 40 years. Their data pertain to current baselines concentrations that would have likely increased over time. The ratios of the trace metal concentrations "near development" and baseline sites reported in the papers by Kelly *et al.* should give an underestimation of the magnitude of the anthropogenic influence.

In contrast, RAMP (2010a) uses the terms *Test* to describe aquatic resources and physical locations (i.e., stations, reaches) downstream of a focal project and *Baseline* to describe aquatic resources and physical locations (i.e., stations, reaches, data) that are (in 2009) or were (prior to 2009) upstream of all focal projects. Data collected from *test* locations are analyzed against *baseline* conditions to assess potential changes. The terms ***test*** and ***baseline*** depend solely on location of the aquatic resource in relation to the location of the focal projects. The Athabasca delta was considered unique in their analysis because there are no regional *baseline* sites that provide a truly adequate comparison. In that case, the *baseline* condition was considered to be all of the previous data from 1998 to 2008. This approach to estimating *baseline* conditions is roughly equivalent to control charting techniques that are designed to determine when processes are out of control and is clearly inappropriate for ascertaining the natural versus anthropogenic inputs of the chemicals of interest. Hebben (2009) did not have baseline stations and hence based his assessment of anthropogenic input on temporal changes (increases) on trace metal (and PAHs) at the monitoring stations. Without reliable data on background concentrations (especially true for metals and PAHs that occur naturally), it would be difficult if not impossible to realistically estimate the relative inputs of the contaminants from natural versus anthropogenic sources.

As well, the laboratories used by RAMP could not achieve the low detection limits needed to measure PAHs, and as a result there were many non-detects. The laboratory used by Kelly *et al.* (2009) was much better at achieving data above detection limits, and therefore they could discriminate concentrations of PAHs among their sampling sites.

2.0 TECHNICAL EVALUATION OF THE REPORTS AND PAPERS

Various limitations were found in all of the four reports reviewed as well as in other documents the WMDRC consulted. The summary below should have important implications for future monitoring work, which could include some re-analyses of older data sets. Specific suggestions for improved data analysis are given in the Appendix.

2.1 General Assessment of the Paper by Kelly *et al.* (2009) on Polycyclic Aromatic Compounds (PACs)

2.1.1 Study Design

In 2008, Kelly *et al.* (2009) conducted a detailed study of the loading of polycyclic aromatic compounds (including, but not limited to the standard suite of polycyclic aromatic hydrocarbons or PAHs) in the Athabasca River catchment. The purpose of their research was to determine the relative contribution of industrial activities to the input of PACs compared with the natural inputs of these compounds from the naturally occurring bitumen. They sampled water (February-March, June-August) using polyethylene membrane devices (PMDs) and snowpack (March).

Sampling sites were chosen in the Athabasca River, its tributaries, the Athabasca Delta and Lake Athabasca. The Athabasca River sites were chosen upstream and downstream of the mining and processing areas; these were in the McMurray geologic formation, which includes the bitumen deposits. It was assumed in the sampling design that the upstream sites would reflect the natural loading of PACs from geological sources only. In four tributary catchments, three sample sites were chosen: one upstream of the McMurray formation and mining areas; one within the McMurray formation but above the mining areas; and one near the confluence of the tributary with the Athabasca River, downstream of the mines and McMurray formation. In the summer sampling, additional downstream sites were included. The choice of upstream sites in the tributary catchments was originally based on 2006 Landsat imagery. However, increasing industrial activity between 2006 and 2008 resulted in loss of some of the tributary sites as reference sites. The change in extent of development between 2006 and 2008 could be evaluated using the Landsat imagery, resulting in the authors categorizing tributary sites as minor or major impacts, each in mid-catchment or at the stream mouth. This separation was used as a basis for subsequent statistical analysis. In addition to the water samples, accumulated snowpack was sampled "at most sites" during the March sampling period.

A concern with the work of Kelly *et al.* (2009) is that it is difficult to identify the sampling locations because of the scale of the map shown in the paper (their Fig. 1) and lack of geographically referenced co-ordinates of the stations (RAMP IT 2010a; 2010b). This is not a consequential criticism but rather a limitation of the PNAS format which restricts paper length. The authors have made detailed site location data available whenever asked. Two major criticisms of the Kelly *et al.* (2009) study include the short period over which samples were collected (less than a year), and the relatively few reference sites. While longer term studies would clearly be advantageous, Kelly *et al.* (2009) were limited by available resources and recognized that their study was, in effect, a short term "pilot" or preliminary analysis of spatial patterns of contaminants adjacent and more remote from the oil sands facilities that could provide baseline data for future measurements rather than a comprehensive study. The few reference sites are, in part, explained by the rapid changes in developed area over only two years (2006 to 2008), making planned reference sites unsuitable after they were chosen. Unfortunately, this is an ongoing fact of life with studies in this area as development continues at a rapid pace. It is essential that this problem of adequate reference sites be addressed in any new long-term monitoring efforts that are undertaken.

Snow samples were collected from 12 sites on the Athabasca River, the Athabasca Delta and Lake Athabasca in March 2008. Duplicate samples were collected at one site on the main river. Snow water equivalents were measured at each site (5 replicates) so that the measured concentrations could be converted to aerial deposition rates for PAC and metals. The snow samples were melted, stirred, sub-sampled, and filtered through 0.45-micron glass fiber filters which were then frozen. The filtrate was spiked with a suite of labeled standards then extracted with a solvent called dichloromethane and stored at -20 degrees C.

PACs in water were measured using polyethylene membrane devices (PMDs), which were placed in the river and tributaries for 30 days, an adequate time for equilibration. The PMDs were rigorously cleaned prior to use. One trip blank and five field blanks were included in both winter and summer sampling periods. There was limited replication of samples; two PMDs were

employed at each site but were located at different depths in the water column. Duplicate PMDs were deployed at two sites, one presumed to have high PAC levels and one presumed to have low levels, but these were located "within 500 m" rather than at the same location. The supplementary information (Kelly *et al.* 2009) appended to the original paper provides extensive details of the sample work-up procedures prior to analysis.

2.1.2 Quality Assurance and Data Validation

The Royal Society of Canada (2010) noted that Kelly *et al.* (2009) sampled at only one location in the river at each site (although at two depths). Kelly *et al.* have subsequently elaborated upon details of the sample collection protocols in written comments submitted to this Committee (E. Kelly, pers. comm. 2011). There is nothing to suggest that the methods they used in sample collection were not scientifically rigorous.

An alternate sampling device, a polyethylene membrane device, is often employed in studies of organic contaminants in water, but as Kelly *et al.* (2009) in correspondence have pointed out, these devices use the compound triolein to absorb the organics. This increases the chance of contamination and also removes the possibility that the sample collected from the membrane device is in equilibrium with the surrounding water.

Samples collected using PMDs necessarily include only the dissolved fraction of the PACs. The snow samples indicate that atmospheric deposition of particulate material containing PACs was a significant input to the landscape. Although transport of the particulate material through the terrestrial portion of the landscape is unlikely to be significant, some of the particulate material is deposited directly on the water and/or ice cover. Kelly *et al.* (2009) measured this portion of the input, at least during the winter period with their snowpack measurements. A comprehensive study should include the particulate component at all times of the year.

Kelly *et al.* (2009) describe analytical methods in detail in the supplementary information attached to their paper. All analyses were conducted at the University of Alberta Biogeochemical Analytical Laboratory using gas chromatography coupled with a mass selective detector. The laboratory at the University has an excellent reputation and has conducted trace organic analyses for a considerable length of time. The detection limits and the reproducibility for the work reported here were excellent, and are well within the values expected of a high quality analytical laboratory. Although concentrations of some analytes were below detection limits, there is no indication that this resulted in any bias in the results or affected the conclusions of their studies. One problem common to studies of PACs is that for some compounds there are no accepted international standards. This applies to all studies of such compounds.

PMDs, filters (from snowpack samples) and geologic samples of oil sands material were extracted, concentrated, loaded on a chromatography column and eluted with pentane to collect the alkane fraction followed by a pentane-dichloromethane mixture to elute the PACs. The extracts of the soluble fraction in the snowpack were filtered through sodium sulphate, concentrated by evaporation and analyzed in hexane. The quality assurance/quality control (QA/QC) procedures included spiking each sample with isotopically-labelled standards. Samples with the chemical signature of diesel fuel (containing only 2 and 3-ring compounds)

were considered contaminated and discarded. This contamination is almost certainly the result of small-scale environmental spills and not a result of the authors sampling techniques. The authors divided the measured compounds into 4 groups: dibenzothiophenes, phenanthrenes/anthracenes, fluoranthenes/pyrenes and benzanthracenes/chrysenes. By comparing the ratios of the classes found in the geologic samples with those found in PMDs and snow, they attempted to identify the sources of the PACs in the samples.
Duplicate samples from sites characterized as "impacted" or "reference" agreed within better than 20%, a level that is acceptable for these chemical parameters. PMD trip and field blanks were low in both winter and summer, although concentrations in field blanks were higher in impacted areas than in reference areas in summer, suggesting that they collected PACs from atmospheric exposure. In some cases in summer, the field blanks yielded higher PAC concentrations than those measured at upstream and midstream sites, indicating that the PACs collected from the atmosphere might have subsequently leached into the water. Although these explanations are entirely plausible, it would have been beneficial to have replication of samples at some of the sites so this could have been assessed more rigourously.

2.1.3 Data Processing and Manipulation (these comments also apply to the trace metals paper by the same authors)

In the absence of high quality emission data (at the time the papers were written) to drive a general atmospheric and deposition model, area-wide deposition of each contaminant was calculated from total mass of the element inferred from snowpack samples in two steps: (i) sample results expressed on a mass per unit area were regressed against distance from Station 6 on the Athabasca River (AR6) assuming that the functional relationship between distance (independent variable) and element mass per unit area was exponential; (ii) assuming that atmospheric flux the total deposition of each contaminant was calculated by finding the integrated deposition within a circle centered on station AR6 and extending to a distance of about 46 km where the data indicated that the deposition was above background.

Kelly *et al.* mostly acknowledged potential problems in estimates of deposition rates. These estimates are very approximate and possibly biased. First, the scatter of points about their regression lines suggests that there are uncertainties in their loading estimates. That, combined with their small sample size, would have generated wide confidence intervals. Second, the deposition rates were calculated with the assumption that wind dispersal of pollutants from oil sands operations would have been about the same in all compass directions. Yet, Kelly *et al.* (2010) acknowledged that winds are predominantly from the north or the south and sampling stations for snowpack were located predominantly north and south of the oil sands operations near station AR6. That situation might have resulted in overestimates of contaminant deposition. On the other hand, snowpack sampling stations were located out in the open on the iced over river and tributaries. If winter winds are strong, those locations might have been subject to wind scour. Contaminant-containing surface snow could have been blown away and redeposited in forests or low-lying areas. That would tend to underestimate the actual deposition rates from the river snowpack data. Further information on these issues may be found in the Appendix.

2.1.4 Conclusions

The quality assurance/quality control (QA/QC) procedures used by Kelly *et al.* would have led to minimal risk of contamination or loss of sample integrity. The Committee concludes that the field methods used by Kelly *et al.* (2009) were adequate in excluding contamination during the sample collection procedures. As well, the laboratory methods employed by Kelly *et al.* (2009) for PAC analysis were well-documented and were satisfactory. There is no indication that their methods caused any erroneous conclusions.

Despite the problems discussed above, the spatial patterns suggested by the Kelly *et al.* (2009) study are noteworthy. The snowpack data provide evidence that PACs are being emitted to the atmosphere by oil sands upgrading facilities and deposited mostly within 30 km of main industrial operations. There are no grounds for concluding that wind-driven transport of bitumen from natural, undisturbed landscapes would be a substantial contributor to PAC deposition in snow pack. Upriver/downriver and disturbed/undisturbed watershed comparisons of concentrations of PAC are also suggestive of input of PAC from oil sands operations, but without additional data on amounts of input of PACs from specific natural sources and seasonal patterns of PAC concentrations in surface waters, little can be said about the relative contributions of natural sources vs. oil sands operations. A more detailed study with more intensive sampling would have been necessary to accurately determine the absolute and relative loadings to the river.

Water collected from the Athabasca River and its tributaries would not reflect the cumulative inputs of contaminants from oil sands operations. Their samples of water and fine sediments from the river represent materials in transit from upriver. Contaminant concentrations in sediments and organisms in the depositional environment of the Athabasca River delta are likely to be among the best indicators of environmental threats. There is moderately strong evidence in Alberta Environment's own data set that PAC levels in delta sediments are gradually increasing over time.

Taking into consideration all data and critiques, the WMDRC agrees with Kelly *et al.*'s conclusion that PACs are being introduced into the environment by oil sands operations. While many of the differences they document are consistent with large inputs from oil sands operations, their water data do not allow even approximate assessment of the relative contributions of natural loadings and those due to oil sands operations.

2.2 General Assessment of Paper by Kelly *et al.* (2010) on Trace Metals

2.2.1 Study Design

Kelly *et al.* (2010) investigated the Athabasca River, its tributaries, the Athabasca Delta, and Lake Athabasca to test the hypothesis that increased concentrations of trace metals in these water bodies are from natural sources, as claimed in the various reports by the Alberta Government (Hebben 2009 and RAMP IT 2010a). Kelly *et al.* (2010) collected winter snowpack samples from 31 sites in March 2008 and surface water samples from 38 sites in February 2008 and 48 sites in June 2008. All sites on the Athabasca River were exposed directly to McMurray Geologic Formation where most of the oil sands occur (this requirement was designed to level out the geologic source contributions) and were chosen upstream or downstream of the oil sands

mining and processing facilities. Three sites along each of four tributaries affected by oil sands development were also sampled along with two undeveloped reference tributaries.

The limitations in the sampling are similar to those discussed above for the PAC paper. In particular, the limited number of samples and the fact that the sampling strategy did not include a time dimension are shortcomings that have been noted in a number of previous commentators on these papers (Royal Society of Canada 2010; RAMP IT 2010b).
The snowpack samples were collected with a plastic shovel, acid-washed Teflon scraper, and an acid-washed Teflon scoop. Samples were placed into acid-washed 2-L Teflon jars (for mercury) or acid-washed wide-mouthed high-density polyethylene bottles (for other trace metals) and stored frozen until analysis. The unfiltered lake/river water samples were collected at all sites using an ultraclean sampling protocol (necessary in any effort to quantify the very low concentrations of trace metals in some of the water and snow samples analyzed). Details of the sample collection protocols have subsequently been elaborated upon by Kelly *et al.* in written comments submitted to this Committee (E. Kelly pers. comm. 2011) and there is nothing to suggest that the methods used in sample collection by Kelly *et al.* (2010) was not scientifically rigorous.

Whether Kelly *et al.* (2010) should have filtered the water samples from the streams, rivers and lake has generated some debate. This problem has remained an enigma and contentious issue in the field of trace metals research since the 1960s. The decision to filter or not filter samples depends on the particular use for the data to be generated. Measurement of total metal concentration would seem logical if the goal of a study is related to source identification, which is the case in the study by Kelly *et al.* (2010). However, the relative distribution of total metal concentrations is generally a poor indicator of the origin of the metals without a clear understanding of the biogeochemical processes in the various parts of the area under study. Instead, differences in metal ratios between reference versus impacted bodies of water would provide a better index of source contributions. The data in Kelly *et al.* (2010) show differences in metal ratios in the tributaries, Athabasca River, Athabasca Delta and Athabasca River. Unfortunately the Committee did not have the time or resources to evaluate the significance of such variations.

Unfiltered (total) snow and river water and filtered (dissolved) snow samples were analyzed for mercury at the University of Alberta Low-Level Mercury Analytical Laboratory by cold vapor atomic fluorescence spectrometry (CVAFS). Samples were analyzed for other elements at the Queen's University Analytical Services Unit using inductively coupled plasma atomic emission spectroscopy with an ultrasonic nebulizer (ICP-AES) and at the Royal Military College Analytical Sciences Group using ICP-MS. Although the two laboratories used in the analysis of metals other than mercury are accredited by the Canadian Association for Laboratory Accreditation to International Organization for Standardization/International Electrotechnical Commission (ISO/IEC) standard 17025, Kelly *et al.* (2010) did not indicate which metals were analyzed by what laboratory or how the results obtained by the two instrumental methods compare with each other.

Samples of bitumen from oil sands were analyzed for trace elements, after microwave digestion, at the Université du Québec á Rimouski Laboratoire de Chimie Marine et Spectrométrie de

Masse, Institut des Sciences de la Mer de Rimouski, by ICP-MS. The three university laboratories used by Kelly *et al.* (2010) are high quality laboratories in the analysis of trace metals using well-established methods and front line instrumentation.

2.2.2 Quality Assurance and Data Validation

Quality assurance must be considered a critical aspect of any monitoring program on trace metals because of the ultra-low levels of the analytes and high susceptibility of samples to contamination. Quality assurance (QA) involves a variety of tasks aimed at preserving the integrity of samples and enhancing the quality of the data and generally includes fieldwork, laboratory analysis, and data validation. Data validation is a process used to determine if data are accurate and complete before it is disseminated prior to its dissemination. The Committee finds the method used by Kelly *et al.* (2010) to be adequate in excluding any overt contamination during sample collection and analysis procedures.

2.2.3 Data Processing and Manipulation

The comments in Section 2.1.3 above on uncertainty of estimated deposition rates are equally valid for trace metals.

2.2.4 General Observations

Kelly *et al.* (2010) showed that 1) the deposition of particulate metals (lead, mercury, nickel and beryllium) decreased sharply with distance from upgrading facilities near Site AR6, similar to the behavior previously demonstrated for polycyclic aromatic compounds (PACs); 2) the deposition patterns for many metals decreased sharply with distance from the upgrading facilities but also increased locally near oil sands development. Metals in this category included particulate antimony, arsenic, cadmium, chromium, copper, silver, thallium, and zinc; and dissolved antimony, chromium, copper, nickel, thallium, and zinc; 3) deposition of particulate lead, mercury, nickel and beryllium in snow was correlated with deposition of particulate polycyclic aromatic compounds (13) ($r^2 > 0.8$, except for mercury, $r^2 = 0.5$; all $P < 0.002$); 4) the concentrations of some trace metals (cadmium, nickel, zinc, mercury, thallium) in tributary water increased significantly near oil sands development and were significantly correlated with overall land disturbance; 5) in winter, concentrations of chromium, mercury, nickel, and silver in the Athabasca River under ice were elevated relative to concentrations just downstream of tailings ponds, impoundments, or other oil sands development infrastructure than upstream, and the concentrations of beryllium, selenium, silver, thallium and zinc were detectable near oil sands development but not upstream; 6) large seasonal differences were observed in levels trace metals in the streams and Athabasca River water samples.

The observations by Kelly *et al.* (2010) (above) provide a body of evidence to show that the oil sands facilities and activities are releasing trace metals into the surrounding environment. The exact amounts and the impacts in the surrounding ecosystems are unclear, however.

Kelly *et al.* (2010) estimated that 34,000 metric tons of airborne particulates were deposited in 2008 within 50 km of upgrading facilities (AR6). The majority of the particulates was said to

consist of oil sands bitumen, as indicated by the large proportion of oil per unit particulate mass and similar distributions of trace metals and polycyclic aromatic compounds in oil sands and particulates. According to the National Pollutant Release Inventory (NPRI 2010), the emission of total particulate matter by oil sands facilities located in all parts of Alberta in 2008 totaled only 7,272 tons. The disparity between the reported emissions (NPRI's value) and deposition (Kelly *et al.*'s value) of particulate matter is huge. The calculated snowpack deposition rates for trace metals near the oil sands facilities are also significant, the total (dissolved + particulate) deposition in March 2008 varying from 1 $\mu g/m^2$ for mercury to 2900 $\mu g/m^2$ for chromium, 3303 $\mu g/m^2$ for nickel and 10,448 $\mu g/m^2$ for zinc. Kelly *et al.* (2010) estimated total snowpack deposition of lead, mercury and nickel over a 4-month in 2008 at sites within a 50-km radius of AR6 Station to be 162, 1.1 and 583 kg, respectively; prorating these values to yearly rates result in annual deposition estimates of 486, 3.3, and 1749 kg/yr, respectively. By contrast, the emissions of total PAC, lead, cadmium and mercury by the entire oil sands industry in 2007/2008 were estimated to be 6/4 kg, 893/761, 134/136, and 82/85 kg (NPRI - National Pollutant Release Inventory, 2010). Assuming that about 20% of these emissions occur in the Athabasca Oil Sands area, based on reported cumulative production in various parts of Alberta (Royal Society of Canada, 2010; Table 2.4), it would appear that the deposition rate for lead (commonly particulate-associated) exceeds the emission rate in the area, whereas only about 20% of the emitted mercury (commonly in gas phase) is deposited locally. The reason for this apparent disparity between the amounts emitted (particulates and many trace metals) and deposited is unclear but strongly indicates a need for an assessment of fugitive sources of airborne particulates in the region. It is also conceivable that considerably more particulate matter and trace metals are being released from the oil sands facilities than is being reported in the NPRI.

As discussed above, Kelly *et al.* (2010) observed an elevated deposition of total mercury in snow pack near areas developed for oil sands production in the Athabasca River region of Alberta, compared to background sites. This pattern led them to suggest that enrichment of mercury deposition occurred locally near oil sands development due upgrading facilities. These observations are interesting and probably should be repeated (see below). However taken at face value they lead to additional questions. There are two Mercury Deposition Network (MDN) sites in Alberta, but further south than the oil sands region (AB 13 Henry Kroeger; AB 14 Genesee). The wet mercury deposition for these sites was 4.3 and 5.9 $\mu g/m^2$-yr in 2007 and 4 and 4.1 $\mu g/m^2$-yr in 2008, respectively. Taken from Figure 1 of Kelly *et al.* (2010), mercury deposition in snow pack for the near development sites appears to be approximately 0.3 $\mu g/m^2$. The background sites appear to be less than these values, approximately 0.05 $\mu g/m^2$. If the accumulated snow pack represents deposition for 3 months, these values could be roughly prorated on an annual basis to 1.2 $\mu g/m^2$-yr and 0.2 $\mu g/m^2$-yr for the near development and background sites, respectively. Although crude, these estimates for atmospheric mercury deposition in oil sand region deposition appear to be well below the regional MDN values. Also it should be noted that the MDN values represent wet-only deposition, while the snow pack should represent total deposition during winter.

This discrepancy is difficult to understand and interpret. If the Kelly *et al.* (2010) estimates are accurate, they would represent very low deposition values. It could be that the Kelly *et al.* (2010) values are simply the result of very limited observations that should be repeated. As a

result of atmospheric emissions and transport of different forms or species, mercury pollution is the result of local, regional and global scale emissions (Evers *et al.* 2007; Driscoll *et al.* 2007). The higher concentrations of mercury in snowpack near development sites reported by Kelly *et al.* (2010) are presumably due to oil sands operations.
Concentrations of total mercury in tributary and Athabasca River water were also measured during winter and summer (Kelly *et al.* 2010). At all sites, they found substantially greater concentrations of mercury during summer than winter. For tributary waters they found significantly greater concentrations of mercury at sites that were less disturbed by development than those that were more disturbed by development. In the Athabasca River, concentrations of mercury were relatively low upstream of the development area. Mercury concentrations were greater in river water immediately downstream of the development area. These concentrations were even less downstream of the development area and in the delta of the Athabasca River. These concentrations of mercury decreased further in Lake Athabasca. It was suggested that contaminants in snowpack are likely released as a pulse to surface waters during spring melt (Kelly *et al.* 2010). This conclusion would only be valid if the mercury was delivered bound to particles from the oil sands operations and would seem unlikely if the emission was in the gas phase since the residence time of gaseous mercury in watersheds is decades to centuries (Demers *et al.* 2007). The range of mercury concentrations reported by Kelly *et al.* (2010) is similar to that reported in RAMP (2009) although the sampling stations are undoubtedly different. It is difficult to make any conclusions of the role of oil sands activities in the supply of riverine mercury in the Athabasca River region without more data.

Findings about concentrations of mercury in biota collected from the oil sands region may be relevant to this discussion Hebert *et al.* (in press). First, mercury concentrations in California Gull eggs from Egg Island (in Lake Athabasca, 60 km NE of the delta) increased 40% between 1977 and 2009 ($P = 0.04$). Second, in 2009 mercury concentrations in Common Tern eggs from Mamawi Lake in the delta area had mercury concentrations 61% higher than in Common Tern eggs from a colony in a more pristine environment (Rocky Point), about 40 km up the Peace River from the delta ($P = 0.0074$). These observations suggest that more attention should be paid to concentrations of contaminants in the biota and sediments of the delta area.

2.3 General Assessment of the Alberta Environment Report by Hebben (2009)

2.3.1 Sampling Design

The rationale behind the long-term monitoring of major lakes and rivers throughout Alberta is to facilitate assessment of provincial regulatory programs, point- and non-point pollution, pollution abatement technologies, the impacts of watershed development activities, and climate change, relative to their impacts on surface water quality. Since stations are generally located both upstream and downstream of major human development areas, the data can be used to support cumulative effects assessment. Samples for the study by Hebben (2009) were obtained monthly from four stations on the Athabasca River: near the headwaters (Hinton), at the town of Athabasca, upstream of Fort McMurray, and 200 km downstream of Fort McMurray at a site called Old Fort. The chemicals monitored at these sites from 1960 to 2007 included trace metals and PAH but with different time frames. The monitoring initiative known as the Long-term River Network (LTRN) program generated the database used in Hebben (2009).

Because of limitations in sampling frequency and number of stations sampled, the LTRN database for trace metals is not adequate to provide significant insights on sources of metals in the river. Measuring inputs from the atmosphere and non-point sources are not part of the LTRN's current mandate. It is unlikely the program could assess the impacts of short-term spikes in trace metal concentrations, such as during spring runoff, and thus it represents a blunt tool for assessing the impacts of provincial regulatory programs on water quality in the oil sands area. Hebben (2009) stated that it was these limitations that forced him to narrow the focus of his report to identifying and describing the long-term changes in water quality parameters. The primary objectives of this report were to (i) provide a general summary of water quality conditions for LTRN sites on the Athabasca River and (ii) examine long-term trends in those data since 1960. An objective was to statistically assess the now-extensive database to lay the groundwork for subsequent investigations into what has changed and why it might have done so. The intention of the report was not to assess potential causes for any of the identified trends, especially for trace metals.

2.3.2 Sample Collection

Hebben (2009) did not provide any information on the field and laboratory procedures and instead referred the reader to the following sources for "further information regarding surface water quality, sampling method and guidelines": Alberta Environment Water Quality Sampling Manual (AENV 2006), the Canadian Environmental Quality Guidelines (CCME 1999) or through the Surface Water Quality homepage (http://www3.gov.ab.ca/env/water/SWQ/index.xfm). Unfortunately, the first two sources had little to say about quality assurance for collecting samples for trace metal analysis and the last source is no longer active.

2.3.3 Laboratory Methods

The samples were analyzed for major ions, nutrient elements, biotic variables, trace metals (Ag, Al, As, B, Ba, Be, Cd, Co, Cr, Cu, Fe, Li, Mn, Mo, Ni, Pb, Sb, Se, Sr, Ti, Tl, V and Zn) and organic compounds including PAHs, extractable priority pollutants, volatile priority pollutants, pulp and paper chlorinated phenols, resin acids and pesticides. This assessment exclusively addresses the data on PAHs and trace metals.

Alberta Environment has developed its own quality assurance procedures for monitoring freshwaters, including quality control samples (Mitchell 2006). As with all sampling programs, it relies on providers to assure the quality of the data it receives. The policy for laboratory data quality assurance in Alberta went into effect in 2001 (Alberta Environment 2004) which requires primary providers of data to the government to be accredited by the Canadian Association for Laboratory Accreditation (CALA).

The reliance on contract laboratories for water quality monitoring raises a number of issues especially with respect to trace substances in natural waters. It would be difficult to validate the data for new metal compounds, which may be of emerging concern where no accredited method or reference standard exists. Laboratory accreditation does not guarantee or ensure good data.

McDonald and LeClair (2004) evaluated quality control data for trace metals from the Alberta Long-term River Network (LTRN) sampling program. They summarized split and spiked samples and evaluated data from two analytical laboratories. The study found that a proportion of trace metal concentrations differed by more than 20% between the two laboratories, and that a number of blank samples contained metals at concentrations above method detection limits. This study draws attention to the need for AENV to provide some oversight on the quality of the work of the contract laboratories.

A scientifically rigorous monitoring program is a science-based approach that uses robust sampling design, consistent methodological techniques, and standardized reporting to generate results that are independent, objective, complete, reliable, verifiable and reproducible. With this definition, the monitoring program used to generate the data reported by Hebben (2009) would appear to be deficient in several respects.

2.3.4 Data Processing/Manipulation

Limitations in the quantity and accuracy of long-term trace metal data made it difficult for Hebben (2009) to apply trend tests to them. A primary limitation pertained to difficulties with measuring the different metal fractions (dissolved, extractable or total) in samples. In response to changes in agency's program interests, enhanced analytical tools and evolving scientific opinion and knowledge, the fractions of metals that were measured changed over the years. Furthermore, different analytical methods were used in measuring the metal fractions, making it difficult to authenticate that the same fraction was being measured consistently and to validate the accuracy of the measurements. Since metal fractions were being quantified, most of the results were below the instrumental detection limits and hence were unsuitable for trend analysis. In addition, most of the early measurements were severely compromised by sample contamination. Older analytical methods were not particularly sensitive to trace metals, making it impossible to compare data obtained using multiple methods. As Hebben (2009) aptly noted, the vast majority of AENV metals data for Athabasca River were not amenable to trend analysis at the time. A large number of the metals were reported as non-detects (or censored data) with the complication that the detection limits for many metals had changed over time. Quality control processes within Alberta Environment (details were not specified) helped to identify a number of metals that had historically shown a tendency towards inaccurate or questionable results. Because of the unreliable nature of the older results, Hebben (2009) limited his analysis to the most recent results (last five years or so).

Incorrect or inappropriate statistical procedures pertaining to Hebben's (2009) report have been listed and explained in the Appendix. These and those of other reports have prompted our recommendation for conducting statistical workshops in the region.

2.3.5 Critical Observations

<u>Trace Metals</u>

Concentrations of many trace metals at the Old Fort sampling location exceeded the guideline values on some occasion (Hebben, 2009). The CCME guideline of 300 µg/L for total iron was exceeded 58% of the time while the 100 µg/L guideline for aluminum was exceeded 51% of the

time. The hardness-based total cadmium guideline was exceeded in 47% of samples that were analyzed. Copper concentrations in 42% of samples exceeded the hardness-based guideline, and the exceedences for total lead and total zinc were about 13% each. Hexavalent chromium exceeded the CCME guideline (1.0 µg/L) in two of 18 samples. The exceedences of many trace metals above the CCME guideline with high frequency at this site is significant and consistent with the observations by Kelly *et al.* (2010).

Although Hebben (2009) maintained that the exceedences were likely due to natural suspended sediments in the lower Athabasca River and its tributaries, he did suggest that "at the same time, however, anthropogenic contributions from both point-(wastewater treatment plant effluents, pulp mill effluents) and non-point sources (resource extraction, forestry, agriculture) cannot be ruled out".

Polycyclic Aromatic Hydrocarbons

The LTRN monitoring of Athabasca River included pesticides, chlorinated phenolic compounds, and other priority pollutants, as well as PAH. Because of limited sampling frequency and since relatively few were above the detection limits, the data on these compounds were deemed inappropriate for trend analysis by Hebben (2009). Instead, only basic summary statistics were presented in the report.

Effort and resources have apparently been devoted to generate the mostly non-detect data for PAH and other organic contaminants. Failure to detect PAHs in the sample cannot be a reflection of available analytical technology between 1990 and 2007 (covered in the report), so it seems likely that concentrations in the river are generally low. The problem most likely relates to the ability of the contract laboratories to measure the low concentrations of PAH found in the water sample. One wonders about the performance standards and the level of quantification of PAH that are acceptable to AENV in its contractual agreements with service providers and whether there is a well developed performance assessment process for providers of analytical support to the LTRN program. The limits of quantification (LOQ) should be defined by the need to discriminate among locations, which is based on thresholds for adverse effects, not by logistic or operational considerations of contract laboratories

2.3.6 Conclusions

Sampling and analysis of trace metals in aquatic ecosystems by Alberta Environment (AENV) has been changing and evolving over time in a manner that makes it impossible to make comparisons between the historic data and currently collected data. Therefore these data are of limited use for achieving the goals of the monitoring program. Too many measurements for trace metals and PAHs are reported as non-detects. In 2004, the government laboratory developed new analytical procedures that allowed detection to the ng/L level for most metals, and as a result, fewer non-detects are being observed. The committee found many problems with statistical analysis also. In most cases these involve issues that are not technically complex. But they are important, as reanalyses might in some cases lead to slightly different conclusions. It is important to provide a rationale for the selection of detection limit values. Non-detect values are only useful if they fall well below the threshold for adverse effects. This may in fact be the case for data collected by Alberta Environment, but the rationale for the selection of appropriate detection limits was not provided to the panel.

2.4 General Assessment of RAMP (2009, 2010)

2.4.1 Sampling Design

The Regional Aquatics Monitoring Program (RAMP) was initiated in 1997 in association with mining development in the Athabasca oil sands region near Fort McMurray, Alberta. RAMP is an industry-funded, multi-stakeholder initiative that monitors aquatic environments in the region. The overall mandate of RAMP is to determine, evaluate, and communicate the state of the aquatic environment and any changes that may result from cumulative resource development within the Regional Municipality of Wood Buffalo. The intent is to integrate aquatic monitoring activities so that long-term trends, regional issues and potential cumulative effects related to oil sands development can be identified and assessed; it is a reactive risk management strategy. Monitoring of atmospheric inputs of pollutants from the oil sands operations is not in RAMP's mandate. In 2009, RAMP focused on six components of boreal aquatic ecosystems: climate and hydrology, water quality, benthic invertebrate communities and sediment quality, fish populations, and acid sensitive lakes. This report focuses on issues in RAMP report that are related to water quality.

The selection of water quality measurement endpoints for RAMP's monitoring program is guided by: (i) water quality measurement endpoints used in the environmental impact assessments of oil sands projects; (ii) a draft list of water quality variables of concern in the lower Athabasca region developed by Cumulative Environmental Management Association (CEMA 2004); (iii) water quality variables of interest listed in the RAMP 5-year report; (iv) results of correlation analysis of the RAMP 1997-2007 water quality dataset indicating significant inter-correlation of various water quality variables, particularly metals; and (v) discussions within the RAMP Technical Program Committee about the importance of various water quality variables to assist in interpreting results of the benthic invertebrate community component and the fish population component, and appropriate analytical strategies for the water quality component. Within this broad scope of effort, the RAMP monitoring program has been concentrated in the Focus Study Area (FSA) defined in the Technical Report as those projects owned and operated by the 2009 industry members of RAMP. An important function of RAMP is to address many of the approval-related monitoring needs for the oil sands industry. Viewed from the priority of work, RAMP monitoring program may be considered to be a service to the oil sands industry to a large extent. Low priority is given to monitoring and reporting on any contaminants that do not fit these criteria. Although a large number of trace metals are apparently measured in the samples collected, only the following metals fit the monitoring criteria and hence are included in the RAMP report: total and dissolved aluminum, total arsenic, total boron, total molybdenum, total strontium and mercury.

2.4.2 Sample Collection

RAMP monitored about 100 water quality parameters at approximately 50 sites in the Athabasca watershed (9 in the Athabasca River) in the fall of each year. In 2009, water quality samples were taken from Athabasca River and Delta at the following locations: upstream of Donald Creek, east and west banks, in winter and fall (*baseline* station); upstream of the Steepbank River, east and west banks, in fall (*test* stations); upstream of the Muskeg River, east and west

banks, in fall (*test* stations); "downstream of development" (near Susan Lake), east and west banks, in winter, spring, summer and fall (*test* stations); and upstream of the Firebag River, cross-channel composite sample, in fall (*test* station). RAMP also attempts to collect three years of seasonal (i.e., winter, spring, summer, fall) *baseline* data from newly established sampling stations before any oil sands development occurs upstream of that station. The frequency of sampling at most stations has been criticized (RAMP Scientific Review 2011) as being insufficient to characterize changes in water quality in the watershed due to the high temporal and spatial variability of water quality parameters in fluvial systems. The determination of atmospheric deposition of contaminants of concern to the watershed is not one of the questions that RAMP has been designed to answer.

2.4.3 Quality Assurance and Data Validation

Details on the RAMP monitoring design and rationale are described in the RAMP Technical Design and Rationale document developed by the RAMP Technical Program Committee (RAMP 2009). Although RAMP has written standard operating procedures (SOP) for field sampling, analytical aspects of the quality assurance (QA) protocols used in the monitoring program are left largely to the contract laboratories. This has led to the use of a mix of analytical approaches. Specific approaches have changed over time as the volume of data has increased. In 2008, RAMP collected over 6,200 water quality observations; from 1997 to 2008, almost 73,000 water quality observations had been collected. The large sample throughput required of the contract laboratories increases the likelihood of errors in data validation. Whether RAMP has written common performance standards for their contract laboratories which specify the quantification limit for each analyte and the acceptable level of quality assurance was not clear; no such document was seen by this committee.

The RAMP (2010a) report does not include information on concentrations of polycyclic aromatic hydrocarbons (PAHs) in the water from the Athabasca area. While RAMP measured concentrations of selected PAHs in surface water from 1997 to 2004, the practice was discontinued since values were consistently less than limits of detection. With the improvement in detection limits, addition of PAHs to future water quality analysis is apparently under consideration (RAMP IT 2010a). The WMDRC encourages RAMP to indeed improve the LOQ values selected for PAHs based on what is ecologically relevant, based on threshold for adverse effects. This should include an assessment of the potential for photo-enhanced toxicity under relevant field conditions.

RAMP (2010a) however reported concentrations of total hydrocarbons in sediment samples collected in 2009 (Figure 2). Their results show that most of the organic carbon consists of heavy oils, asphalts, and many PAHs of petrogenic or biogenic origin.

Of greater significance are temporal trends in PAH levels in delta sediments. RAMP (2010a) reported PAH levels for 2000-2009 in sediments for four stations in the Athabasca River delta, noted some high values for the 2009 data set, claimed these values have meaning only when normalized to total organic carbon (TOC), and concluded, without statistical analysis, that there are "no consistent trends over time" (p. iii). Those conclusions were accepted by the recent Royal Society of Canada report (Royal Society of Canada 2010, p. 145). It states (p. 145):

"More recent data (including 2008 and 2009 results) confirm only the variability, but do not show any clear upward trend for alkylated PAHs in sediments for the Athabasca Delta (RAMP 2010). When data back to 2001 are compared on the basis of being normalized to 1% TOC, to account for variability in organic content of sediment samples, an adjustment that is scientifically sound based on the behaviour of PAHs in water in contact with sediment, there is no upward trend in PAH at any of the four sites in the Athabasca Delta" (RAMP 2010).

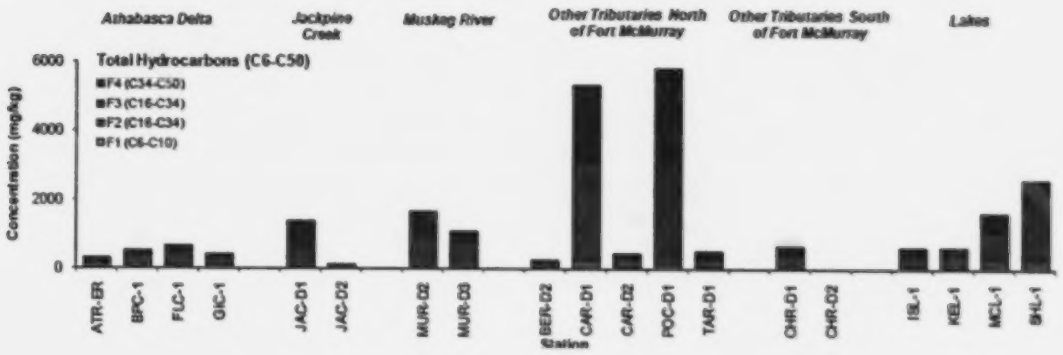


Figure 2 Total hydrocarbon in sediments collected by RAMP in 2009
Note that Fractions F3 and F4 (representing most of the organic carbon contents), with more than 16 carbon atoms, contain heavy oils, asphalts, and many PAHs of petrogenic or biogenic origin. Figure 7.3-6 from RAMP (2010a)

Timoney and Lee (2009) had earlier analyzed the 2000-2007 portion of this data set and shown through regression analysis that there was a marked upward trend in the absolute concentrations of PAH in delta sediments. That conclusion will be reinforced when they repeat their analysis with the 2008-2009 data added in (as they are doing). For three of the four delta stations, the highest observed sediment PAH levels during the 2000-2009 period occurred during 2008 or 2009 (RAMP 2010, pp. 5-49 - 5-52). We fault some aspects of the statistical procedures used by Timoney and Lee (see Appendix), but their correction is unlikely to require any change in their conclusions.

The WMDRC disagrees with the assumption of RAMP (2010a) and Royal Society of Canada (2010) that the TOC-normalized PAH levels is the most or only relevant index of PAH contamination. Such normalization is rarely done in contaminant-ecosystem studies. One can envisage various causes of increased sediment TOC over the last decade, e.g. increased algal and macrophyte production due to increased, nutrient-bearing inflows of wastewater inflows from the growing human population in the region. Between 1999 and 2009, the population of Alberta increased by 25%.

But we do not see how increasing organic matter from such sources -- or other sources or causes -- can justifiably be used to minimize the significance of increased absolute PAH levels in delta sediments, as RAMP (2010a) and the Royal Society of Canada Report (2010) attempt to do.

Effects on the sediment biota, including microbes, filter- and deposit-feeding invertebrates, and bottom-feeding fish -- and organisms that may feed on them, such as birds and mammals, including humans -- are likely to relate directly to the absolute concentrations (mg/kg of dry weight sediment) of PAHs in the sediments. If increased PAH concentrations in sediments are thought to reflect phenomena, e.g. climate change, reduced flows, increased domestic wastewater inputs, etc., other than oil sands operations, then more sophisticated analytical approaches need to be applied to these data sets.

RAMP also reported a parameter for sediments called "total metals" said to vary from less than 30 mg/kg to nearly 500 mg/kg. Metals included in this parameter were As, Ba, Be, Cd, Cr, Co, Cu, Pb, Hg, Mo, Ni, Se, Ag, U, V and Zn. How this aggregate measure was derived was not explained in the RAMP report – whether based on absolute concentration, equivalent concentrations, toxic equivalent or other functional relationships. We feel that the term "total metals" is meaningless, unless done on a toxic unit approach where an index of total metal toxicity equivalents is determined. This could be done for some metals with similar modes of action and for which there is no or little non-additivity of toxicity. However, this cannot be done for all of the metals listed by RAMP.

It is notable that over the past decade (RAMP IT 2010a) found no detectable trend for arsenic concentrations in Athabasca delta sediments. In contrast, between 1970 and 1990, arsenic concentrations increased 5-fold in the sediments of Lake Athabasca (Bourbonniere *et al.* 1996).

Total metal values normalized to percentage concentrations of silt and clay were also reported. The normalization seems odd considering the finding by RAMP that the total metal concentrations are correlated with fine sediment fractions in the RAMP sediments database. From 2006 to2009, they showed from principal component analysis that "total metals" was strongly and positively correlated with % silt (rs=0.87) and % clay (rs=0.77). If clays and silts are predictors for total metals, normalization with such covariates may be used to identify long-term (decadal) changes rather than the year-to-year trend.

2.4.4 Data Processing and Manipulation

The water quality component of RAMP is currently intended to address four specific null hypotheses (RAMP, 2009): 1) Water quality at each sampled location is within the range of natural or baseline variability (H 1); 2) Water quality at sampled locations does not change over time (H 2); 3) Water quality at upstream and downstream sampling locations is similar (H 3); and 4. Water quality characteristics at each sampling location do not exceed relevant environmental quality guidelines; (H 4). H 1 is tested using a regional reference approach, whereby values are compared to the range of regional natural or baseline variability (RAMP, 2009). H 2 and H 3 are not assessed statistically due to the inherent variability of water quality data, but are assessed using semi-quantitative trend analysis. Data are compared to water quality guidelines (AENV, 1999; CCME, 2002) to test H 4. Comments on the suitability of these approaches are also contained in the 2010 RAMP external review (RAMP External Review 2011).

The WMDRC has concluded that the Water Quality Index (WQI) is not scientifically rigorous and may foster superficial interpretations and summaries. The construction and weak rationale for it are given in RAMP (2009, p. 3-63). What decision-makers and the general public need from scientists is clear description of the problems, possibly presented by trends in the key individual measures of water quality or contamination threat, taken one by one. Monitoring and science seems to be ill-served by the WQI as used to convey the water quality status in the oil sands area. Valuable information is lost in constructing the WQI, which makes it more difficult to discern the status and trends.

Data analysis by RAMP has many limitations described in detail in the Appendix.

2.4.5 Conclusions

The RAMP data (particularly with respect to metals) are not directly comparable to the data from Kelly *et. al.* (2009, 2010). The majority of the RAMP data is from fall sampling, as opposed to the winter and spring sampling, and it is difficult to determine if the sampling sites are directly comparable. Overall the data presented in Hebben (2009) for the AENV stations "upstream of Fort McMurray" and "Old Fort" are more pertinent to the questions posed to the committee, since the sampling is monthly, as opposed to once a year in the fall.

Despite the limitations noted above, the Regional Synthesis (Section 7) in the 2009 RAMP Report (RAMP IT 2010a) provides some insight to the question of change in the Athabasca watershed. In 2009, the concentrations of total aluminum in the Athabasca designated as test were generally greater than those designated as baseline. While there were frequent instances of total aluminum concentration exceeding water quality guidelines, dissolved aluminum concentrations (a better measure of metal bioavailability and hence potential impact on aquatic organisms) were normally below guidelines. While there were no indications of year over year increases in total and dissolved iron concentrations in the watershed, there were numerous instances where both the total and dissolved concentrations exceeded applicable water quality guidelines. Finally and perhaps most significant, there was "a general increase in frequency of measurable concentrations of mercury among all baseline and test stations monitored by RAMP" (RAMP IT 2010a).

In response to the Kelly *et al.* (2009, 2010) papers, RAMP Scientific Review (2011) assessed whether there were correlations between the dissolved and total metal concentrations in these lakes and distances of the lakes from major upgrading facilities (AR6). They did this on the assumption that "if aerial deposition was continually contributing metals to surface waters the distribution of metals in these lakes should reflect the levels of deposition to which each catchment is exposed" (their p.3). They found "no spatial or temporal trends in regional ponds or lakes that would suggest a measurable effect of atmospheric deposition on lake water quality" (their p. 8), though presenting only a few scatterplots supporting that conclusion. That critique should not be interpreted as negating, by itself, any of Kelly *et al.*'s snowpack-derived estimates of deposition rates. The underlying assumption of RAMP IT (2010a) is not reasonable. These 50 lakes differ from each other in their morphology, hydrology, water chemistry, and food webs. The behavior of contaminants in them will be equally variable from lake to lake. If instantaneous contaminant deposition rates were the same on all lakes, a very short time later one

would expect to find large differences among the lakes in levels of the contaminant in their water columns. Lakes represent a very dynamic biogeochemical environment compared to that of the snowpack and are completely unsuited for assessing, even approximately, deposition rates.

3.0 GENERAL CONCLUSIONS

The committee evaluated four documents on the impacts of the oil sands on water quality in the Athabasca River system. Note that there are other monitoring programs investigating environmental impacts in this area. Our focus, however, was mainly on the two papers by Kelly *et al.* (2009, 2010), one Alberta Environment report and one RAMP report.

Chemical monitoring in fluvial systems, particularly large complex systems such as the Athabasca River and its tributaries, is not easy. Concentrations can vary substantially over time (season, flow rate) and over space within the system. It is difficult to obtain a database that is sufficiently large to make any meaningful assessment of environmental risk of the concentrations of chemicals in the system. This is especially true when those concentrations are at or close to the threshold of concentrations known to have an effect.

The focused, short-term sampling campaign used by Kelly *et al.* (2009) was adequate for assessing short-term inputs to the aquatic ecosystem. The Alberta Environment monitoring at limited number of stations was not specifically intended to determine impacts from the oil sands operations, but is valuable in trend analysis. Unfortunately, trends in PAH concentrations over time in the Athabasca River cannot be ascertained because for almost all water samples analyzed, PAH levels were below analytical detection limits. The RAMP program has many monitoring sites, but the low sampling frequency limits their ability to determine impacts from oil sands operations.

There are caveats in using concentrations of contaminants in river water and sediments to assess environmental impacts. The water and sediments in a stream or river represent materials in transit and will always be dominated, at any given location, by materials more or less recently arrived from upriver. Even with constant inputs from oil sands operations one would not necessarily expect an increase over time in concentrations in water and sediment in rivers. Fluxes or loadings should be calculated that give the mass of a contaminant moving from different tributaries to points downriver.

Kelly *et al.*'s. (2009, 2010) articles and the Alberta Environment document are based on monitoring chemicals that are predicted to exceed a threshold for effects based on controlled laboratory studies, as opposed to actual concentrations in biota and actual effects on organisms (biological monitoring). No biological data on how the chemicals affect organisms are presented in any of the material. These documents infer the potential for biological impact solely by comparing chemical concentrations to either water quality guidelines (for trace metals) or published concentrations that may cause an effect to certain aquatic organisms (for PAHs) – Tier 1 risk assessment approach.

The Kelly *et al.* (2009, 2010) papers have served a useful purpose in pointing out some deficiencies in the current monitoring programs in the oil sands area. The committee thinks that

this is now accepted by all, as evidenced by the Alberta Premier appointing yet another panel to develop a more rigorous monitoring program for the future. We believe that this was the goal of Prof. Schindler's team when they initiated their studies. They did what they could do with limited funds, and while not perfect and limited in scope, the results have received considerable attention.

It seems likely there are limitations on the ability of Kelly *et al.* (2009, 2010) ability to estimate the mass loading of contaminants from the oil sands activities into the watershed. Their results, however, carry the implication that considerably more particulate matter and trace metals are being released from the oil sands facilities than is being reported in the National Pollutant Release Inventory or that fugitive sources are a major source of local pollution in the area. Regardless of the reasons for the large disparities between estimated emissions from oil sands operations and estimated depositions of pollutants around these facilities, the papers by Kelly *et al.* (2009; 2010) have served to point out some of the gaps in data and understanding in linking sources and effects.

Limitations in the quantity and accuracy of long-term trace metal data made it difficult for Hebben (2009) to apply trend tests to them. A primary limitation pertained to difficulties with measuring the different metal fractions (dissolved, extractable or total) in the samples. In response to changes in the agency's program interests, enhanced analytical tools and evolving scientific opinion and knowledge, the fractions of metals that were measured were changed over the years. Furthermore, different analytical methods were used in measuring the metal fractions, making it difficult to authenticate that the same fraction was being measured consistently and to validate the accuracy of the measurements over the long term. Since metal fractions were being quantified, most of the results were below the instrumental detection limits and hence were unsuitable for trend analysis. In addition, most of the early measurements were severely compromised by sample contamination.

Hebben (2009) reported that concentrations of many trace metals at the Athabasca and Old Fort sampling locations in the Athabasca River exceeded the guideline values on many occasions. The Canadian Council of Ministers of the Environment (CCME) guideline of 300 µg/L for total iron was exceeded 58% of the time while the 100 µg/L guideline for aluminum was exceeded 51% of the time. The hardness-based total cadmium guideline was exceeded in 47% of samples that were analyzed. Copper concentrations in 42% of samples exceeded the hardness-based guideline, and the exceedences for total lead and total zinc were about 13% each. Hexavalent chromium exceeded the Canadian Council of Ministers of the Environment (CCME) guideline (1.0 µg/L) in two of 18 samples. The exceedences of some trace metals above the CCME guideline at this site with great frequency is significant and consistent with the observations by Kelly *et al.* (2010).

The LTRN monitoring of the Athabasca River included PAH and other organic contaminants, such as pesticides, chlorinated phenolic compounds, and other priority pollutants. Because of limited sampling frequency and since relatively few were above the detection limits, the data on these compounds were deemed inappropriate for trend analysis by Hebben (2009). Instead, only basic summary statistics were presented in the report.

Significant effort and resources have apparently been devoted to generate the mostly non-detect data for PAH and other organic contaminants in LTRN monitoring. Failure to detect PAHs in the sample cannot be a reflection of available analytical technology between 1990 and 2007 (covered in the report), so it seems likely that concentrations in the river are generally low. This most likely relates to the inability of the contract laboratories to measure the low concentrations of PAH found in the water sample.

A scientifically rigorous monitoring program is a science-based approach that uses robust sampling design, consistent methodological techniques, and standardized reporting to generate results that are independent, objective, complete, reliable, verifiable and reproducible. Viewed through this prism, the monitoring program used to generate the data reported by Hebben (2009) would appear to be deficient in several respects.

RAMP is spending time and resources on obtaining water quality data that are difficult to interpret because the systems they are monitoring are large, complex and variable, and their sampling frequency is too low and stations are not well situated to account for this.

Although a reasonable range of trace metals are apparently measured in the samples collected, only the following metals fit the monitoring criteria and hence are included in the RAMP report: total and dissolved aluminum, total arsenic, total boron, total molybdenum, total strontium and mercury. Data for all trace metals measured in water samples should be included in future reports by RAMP.

RAMP (2010a) measured concentrations of total hydrocarbons in sediments collected in 2009. Their results show that the principal components of organic carbon (over 95%) were heavy oils, asphalts, and many PACs of petrogenic or biogenic origin.

Conclusions in the two Kelly *et al.* papers and the AENV report reviewed by the committee are based on monitoring chemicals that may cause an effect, as opposed to monitoring actual effects on organisms (biological monitoring). No biological data on how the chemicals affect organisms are presented in these documents. These reports attempted to infer the potential for biological impact by comparing chemical concentrations in their samples to either water quality guidelines (for trace metals) or published concentrations that may cause an effect to certain aquatic organisms (for polycyclic aromatic compounds, PACs). The need for data that can be used in comprehensive risk assessment cannot be over-emphasized.

The current structure and operation of RAMP are such that the program is regarded with suspicion by some members of the community. The fact that the data and their interpretations were not released to the public or available to interested parties severely damaged the credibility of the RAMP program. While the consultants employed to do the monitoring are capable, the involvement of an independent steering committee to provide oversight, encourage collateral science-based research and help to ensure coordination of the water (RAMP) and atmospheric (WBEA) monitoring programs (such as co-locating the monitoring stations) may help increase the credibility of RAMP programs. The independent steering committee should report directly to a provincial minister.

Assessing the contributions of oil sands operations to contaminant levels is made difficult by the scarcity of information on natural historical background levels and on true reference sites. This is generally understood, but must be kept in mind where the various reports refer to "upstream" versus "downstream," "background" versus "near development," and "test" versus "baseline" sites or conditions. There are at least four underlying causes for absence of adequate reference sites. First, no strong monitoring programs were put in place until well after oil sands operations began. Second, once they were in place occasional changes in analytical labs used and in detection limits have complicated assessment of temporal trends. Third, locations and amounts of natural inputs, especially by groundwater inflows, are poorly known but in many cases are upriver of oil sands operations. And fourth, aerial dispersal and subsequent deposition of contaminants generated by oil sands operations will often be southward, e.g. upriver of Fort McMurray, as winds are frequently from the north.

Given this situation, other techniques need to be employed to assess anthropogenic burdens in the local environment and biota – such as isotopic fingerprinting methods, metal ratios and enrichment factors, changes in relative ratios of congeners and homologues of organic compounds. Temporal trends in concentrations are very susceptible to changes in climate and hence are not very appropriate for ascertaining the industrial component of a contaminant in local lakes and rivers.

There are several reasons for the apparent differences of opinion ("the controversy"), the two main ones being (A) each of the databases is limited in terms of quality, quantity and/or lack of spatial and temporal resolutions and cannot scientifically justify the inferences that have been reported and (B) the use of different reference sites and hence benchmark concentrations to estimate the burden of contaminants from anthropogenic sources.

In the end it is not really a quest of "who is right", but rather, what questions were they asking and how relevant were the data collected by each group relative to their different purposes and hypotheses to be tested. Here, we have evaluated the sufficiency of the data collected by various groups to meet the needs of their individual mandates and purposes. Each of the studies presents some useful information and each suffers from some limitations.

Subsequent to the studies by Kelly et al., the Alberta government has begun some studies of deposition and included some elements of the air-monitoring program just recently set up by the Wood Buffalo Environmental Association. These new initiatives will be addressing some of the deficiencies noted by Kelly et al. Alberta Environment has also launched a contaminant study in the area.

4.0 RECOMMENDATIONS

In the course of reviewing more than a thousand pages of documents and interviewing 18 people involved in monitoring and scientific work relating to regional oil sands operations, the WMDRC reached a number of general conclusions about steps that would improve the quality of both the monitoring and scientific rigor of the water quality monitoring programs. We give these here fully realizing that other groups have been and will be preparing reviews on these same issues.

1. Monitoring programs should be reoriented to focus on documenting status and trends in key pollutants over time and space, including contaminant deposition and loading rates, and actual concentrations. More emphasis should be given to generate data to improve our understanding of the distribution and risks of contaminants, especially those that can be used as biomarkers and bioindicators of oil sands pollution (such as specific congeners of organic PAH and other compounds; BTEX – benzene, toluene, ethylbenzene and xylene; glycol; vanadium, titanium, zirconium, mercury, arsenic, selenium, thallium, etc) and effects of such contaminants. Such redefinition of main objectives should guide any future changes in the sampling designs and monitoring activities in the field. The monitoring should be designed to answer the questions: Are the right variables and the most sensitive or informative ones, being measured in the right places? Occasional and well-timed intensive monitoring should be conducted to capture short-term pulses in contaminant concentrations that may not be obvious in weekly or monthly samples. Monitoring of tributaries should be given more attention, especially at critical periods of low flow conditions and reproduction of key species.

2. More effort and resources need to be put into making the monitoring program more scientifically rigorous, into ensuring that the data are collected in an unbiased manner and into making the database available in the peer-review scientific literature.

3. Effort should be made to build and expand local capacity on statistical analysis of large and multi-dimensional databases. Statistics workshops of a refresher nature should be established and funding provided to bring together perhaps three to four dozen persons active in oil sands research, monitoring, data analysis, and writing. They will have enough expertise among themselves to sort through the issues raised here, but having some experienced professional statisticians from the outside would also be useful. Formal involvement of the Statistical Society of Canada might be appropriate.

4. Detailed year-around and multi-year measurements of air mercury species and wet deposition of mercury and other toxic metals (similar to the Mercury Deposition Network) should be implemented for the oil sands region to provide improved characterization of local sources of mercury emissions and mercury deposition. The occurrence of elevated measurements of atmospheric mercury and toxic metal concentrations or depositions at these sites could be evaluated by back trajectory analysis to determine the source area. This program may be part of the plans for the Wood Buffalo Environmental Association atmospheric chemistry and deposition program. If so, this program should be closely coordinated (by co-locating the sampling sites for instance) and hence able to provide additional insight on mercury and other toxic metal

deposition to the region and the contribution of the oil sands activities to the atmospheric flux.

5. A detailed survey of trace metals and PAHs in snow should be conducted to repeat the observations of Kelly *et al.* (2009, 2010). However this snow survey should involve more transects away from the oil sands development site and across the diverse landscape to better characterize the transport and deposition of mercury from the site.

6. A paleolimnological study should be initiated to evaluate historical trace metals, mercury and PAC deposition in lake and bog sediments from undisturbed watersheds proximate and remote from the oil sands development area. Sediment cores should be collected from several lakes both near and remote from the oil sands area. These cores should be age dated (by ^{210}Pb or other geochronological methods) and measured for mercury and other relevant contaminants. Profiles of stable metal isotopes, metal ratios, and changes in carbon contents of various organic compounds can provide information on historical deposition to the region and help quantify the influence of emissions from oil sands activities on contaminant burdens in sediments.

7. The contract laboratories should develop methods for trace metals, PAHs and other organic contaminant analyses that are sensitive enough to determine the actual concentrations of contaminants in environmental media and reduce the number of non-detects being reported. Alberta Environment should endeavour to establish clear performance standards (or tighten them up if they exist already) that stipulate the level of quantification of PAH and trace metals especially in water samples, and establish a performance assessment process for providers of analytical support to the LTRN program.

8. Measurements of trace metals should include total and filtered forms, and methyl mercury, coincident with measurements of dissolved organic carbon and particulate organic carbon (or total suspended solids). These measurements would provide a better understanding of the characteristics of surface water trace contaminants. These fluvial measurements could be coupled with discharge measurements to calculate the mass flux of trace contaminants. These fluxes should be compared to estimates of atmospheric deposition of trace contaminants.

9. Polycyclic Aromatic Hydrocarbons (PAH) can exhibit photo-enhanced toxicity that can make these compounds thousands of times more toxic in natural environments in the presence of solar radiation (Newsted and Giesy, 1987, Oris *et al.*, 1986, 1987). For this reason, the WMDRC suggests that this phenomenon be considered in interpreting the potential effects of PAH in both terrestrial and aquatic environments.

10. More attention should be given to naphthenic acids (NA) in the Alberta Environment and RAMP monitoring programs. The NAs are a large class of cyclic, organic compounds that are associated with oil sands. Much of the toxicity of oil sands process water (OSPW) has been attributed to NAs so they should be considered in future monitoring programs. This is a complex mixture that is difficult to characterize, but progress is being made in their analytical characterization and understanding of their toxicity. Some specific biomarkers could be developed to monitor for the effects of these compounds.

11. The RAMP program has some positive elements and a great deal of information is being collected. The information collected, while primarily to meet regulatory requirements could be used in a wider context. Not making the information collected by RAMP available to interested persons has resulted in the credibility of RAMP being impugned. While this situation is being rectified by making the data available on a web site, the damage to the credibility had already been done. To address this issue, RAMP needs a policy of being more open and more rigorous in the interpretation of the data collected. While the consultants employed are capable, it is suggested that an independent panel of scientists be retained to advise and help interpret the results. These individuals should be of the highest technical credibility and have credentials beyond reproach, such as being members of National Academies. To maintain accountability, the panel members should be remunerated much as is done by Environment Canada or the US Environmental Protection Agency and they should be appointed by and serve at the pleasure of the Alberta Minister of Environment. This is a proven model that will produce the best quality science and maintain the credibility an important program like RAMP needs in order to be effective. The current structure is totally unacceptable.

12. Participation of local communities is absolutely essential to the long-term viability of the exploitation of the oil sands as a resource. Every effort should be made to involve and communicate with all interested parties and stakeholders.

13. The oil sands are an important national resource and as such, Environment Canada (EC) and Health Canada should be more engaged in the studies being conducted. Of course, they would need to be closely integrated with and provincial efforts. But EC has unique capabilities, expertise and equipment that can be brought to bear on environmental issues involving the oil sands and their exploitation. EC in conjunction with academic researchers could best conduct targeted studies to answer specific questions as opposed to long-term monitoring that can best be done by the province, RAMP and Wood Buffalo Environment Association. The WMDRC understands that some of these sorts of efforts are in the planning stages and endorses this activity.

14. Across Canada, the Province of Alberta has some of the very best environmental scientists in the world. The WMDRC encourages their involvement in ongoing planning and management of research and monitoring programs.

5.0 LITERATURE CITED

Alberta Energy. 2010. http://www.energy.alberta.ca/OurBusiness/oilsands.asp

Alberta Environment. 2004. Alberta Environment Laboratory Data Quality Assurance Policy Procedures and Guidelines. http://environment.gov.ab.ca/info/library/6995.pdf

Alberta Environment. 2006. Aquatic ecosystems field sampling protocols. Pub. W0605. 149 p.

Bourbonniere, R.A, S.L. Telford and J.B. Kemper. 1996. Depositional history of sediments in Lake Athabasca: geochronology, bulk parameters, contaminants and biogeochemical markers. Project Report 72, Northern River Basins Study, Edmonton.

Brua, R,. K. Cash and J. Culp. 2003. Assessment of natural and anthropogenic impacts of oil sands contaminants within the northern river basins – Final summary report, Task 5: Hydrocarbons/oil sands and heavy oil research and development. Panel on Energy Research and Development. Reproduced in NREI: Collective findings, compiled by F. Conly, Environment Canada, Saskatoon 2004.

CCME. 1999. Canadian Environmental Quality Guidelines. Canadian Council of Ministers of the Environment, Winnipeg.

Cumulative Environmental Management Association (CEMA). 2004. Development of reach specific water quality objectives for variables of concern in the lower Athabasca River: Identification of variables of concern and assessment of the adequacy of available guidelines.

Demers, J.D., C.T. Driscoll, T.J. Fahey, and J.B. Yavitt. 2007. Mercury cycling in litter and soil in different forest types in the Adirondack region, New York, USA. Ecol. Appl. 17(5):1341-1351.

Driscoll, C.T., Y-J. Han, C.Y. Chen, D.C. Evers, K.F. Lambert, T.M. Holsen, N.C. Kamman, and R.K. Munson. 2007. Mercury contamination in forest and freshwater ecosystems in the Northeastern United States. BioScience 57:17-28.

Energy Resources Conservation Board. 2010. http://www.ercb.ca/

Environment Canada, Alberta Environment, Govt. Northwest Territories. 2004. Northern Rivers Ecosystem Initiative. Key Findings; Synthesis Report. http://www.ec.gc.ca/Publications/default.asp?lang=En&xml=242DCF56-D95A-415B-81ED-D1CA51086853

Evers, D.C., Y-J. Han, C.T. Driscoll, N.C. Kamman, W.M. Goodale, K.F. Lambert, T.M. Holsen, C.Y. Chen, T.A. Clair and T. Butler. 2007. Biological mercury hotspots in the Northeastern United States and Southeastern Canada. BioScience 57:1-15.

Hatfield Consultants. 2009. Technical design and rationale. Prep. for RAMP Steering Committee. RAMP 1467.1. 348 p.

Hebben, T. 2009. Analysis of water quality conditions and trends for the Long-term River Network: Athabasca River, 1960-2007. Alberta Environment, Water Policy Branch, Environmental Assurance. 341 p.

Hebert, C.E., D.V. Chip, S. MacMillan, D. Campbell and W. Nordstrom. 2011. Metals and polycyclic aromatic hydrocarbons in colonial water bird eggs from Lake Athabasca and the Peace-Athabasca delta, Canada. Env. Toxicology and Chemistry (in press).

Helsel, D.R. 2006. Fabricating data: how substituting values for nondetects can ruin results, and what can be done about it. Chemosphere 65:2434-2439.

Kelly, E.N., J.W. Short, D.W. Schindler, P.V. Hodson, M. Ma, A.K. Kwan, B.L. Fortin. 2009. Oil sands development contributes polycyclic aromatic compounds to the Athabasca River and its tributaries. Proceedings of the National Academy of Sciences, 106:52, 22346-22351.

Kelly, E.N., D.W. Schindler, P.V. Hodson, J.W. Short, R. Radmanovich, C.C. Nielsen. 2010. Oil sand development contributes elements toxic at low concentrations to the Athabasca River and its tributaries. Proceedings of the National Academy of Sciences, 107:37, 16178-16183.

McDonald, D. and D. LeClair. 2004. Methods and quality assurance investigations for trace metals data from the Long-Term River Network, 2003. Env. Monit. and Eval. Br. Alberta Environment. 77 p.

Mitchell, P. 2006. Guidelines for quality assurance and quality control in surface water quality programs in Alberta. Prep. for Alberta Environment, Pub. WO603. Environmental Monitoring and Evaluation Branch. 57 p.

National Pollutant Release Inventory. 2010.

Newsted, J.L. and J.P. Giesy. 1987. Predictive models for photo-induced acute toxicity of polycyclic aromatic hydrocarbons to *Daphnia magna* Strauss (Cladocera: Crustacea). *Environ. Toxicol. Chem.* 6:445-461.

Oris, J.T. and J.P. Giesy. 1986. Photo-induced toxicity of anthracene to juvenile bluegill sunfish (*Lepomis machrochirus* Rafinesque.): Photoperiod Effects and Predictive Hazard Evaluation. *Environ. Toxicol. Chem.* 5:761-768.

Oris, J.T. and J.P. Giesy. 1987. The Photo-induced toxicity of polycyclic aromatic hydrocarbons to larvae of the fathead minnow (*Pimephales promelas*). *Chemosphere*. 16:1395-1404.

RAMP (Regional aquatics monitoring program). 2010. http://www.ramp-alberta.org/RAMP.aspx

RAMPIT (RAMP Implementation Team) 2010a. Regional Aquatics Monitoring Program 2009 Technical Report. Prep. for RAMP Steering Committee. 803 p.

RAMPIT 2010b. RAMP Implementation Team unpublished review of Kelly *et al.* papers.

RAMP Scientific Review. 2011. Alberta Innovates – Technology Futures, Calgary. 160 p.

Regional Aquatics Monitoring Program, Wood Buffalo Environmental Association, Cumulative Environmental Management Association. 2008. Joint community update 2008. 24 p.

Royal Society of Canada (RSC). 2010. Environmental and health impacts of Canada's oil sands industry. Royal Society of Canada Expert Panel report. 414 p.

Timoney, K. and P. Lee. 2009. Does the tar sands industry pollute? The scientific evidence. The Open Conservation Biology Journal, 2009, 3, 65-81.

6.0 GLOSSARY

Anthropogenic: human-caused or related to humans.

Aquatic ecosystem: any water environment, from small to large, from pond to ocean, in which plants and animals interact with the chemical and physical features of the water body.

Background, baseline, reference: although there are subtle differences in the meanings of these terms, they generally refer to unimpacted sites that can be compared with impacted sites.

Bioavailability: the degree and rate at which a substance is absorbed into a living system or is made available at the site of physiological activity.

Biogenic: produced by living organisms or biological processes; necessary for the maintenance of life processes.

Biological monitoring: monitoring of biota in the environment. Biological monitoring may include vegetative surveys, phytoplankton, zooplankton or invertebrate analyses, fisheries surveys, or rare and endangered species inventories.

Bitumen: a naturally occurring, viscous mixture of hydrocarbons that contains high levels of sulphur and nitrogen compounds. Bitumen makes up about 10% of the oil sands.

Blank sample: a clean sample or sample of matrix processed so as to measure artifacts in the sampling and analysis process.

Chemical: the makeup of all matter. Water is a chemical, and all life is composed of organic chemicals.

Concentration: the amount of a substance in a given amount of water. For example, milligrams of substance in a litre of water.

Contaminants: substances in water, air, or soil that are not normally present. Usually used for substances of concern for aquatic or human health, although sometimes includes naturally occurring substances.

Cumulative: growing in quantity, strength, or effect by successive additions or gradual steps.

Deposition, atmospheric deposition: chemicals transported by air currents and deposited on the ground or in water bodies.

Detection, detection limit: refers to the lowest concentration of a substance that analytical instruments can detect. "Undetectable" does not mean the substance isn't there, but that the instruments aren't capable of getting a measurement of it.

Emissions: gases and particulates given off by industrial activities, which may include a variety of harmful chemicals.

In-situ extraction: Latin for "in place". In situ extraction refers to various methods used to recover deeply buried bitumen deposits, usually by pumping steam down to them so they will be soft enough to draw up the bitumen.

Landsat: any of various satellites operated by U.S. government organizations, used to gather data for constructing images of the Earth's surface.

Loading: quantities of substances that enter a water body or the air over a given time period.

Microwave digestion: the assisted solubilisation of a sample by the application of radiation.

Model, modeling: computer program used to describe and make predictions about environmental events.

Monitoring: a sampling program designed to document natural events or impacts in the environment.

Non-point source: diffuse or undefined sources that are usually carried by runoff.

Oil sands: bitumen-soaked sand, located in four geographic regions of Alberta: Athabasca, Wabasca, Cold Lake, and Peace River. Bitumen can be extracted either by surface mining or in-situ.

PAC (Polycyclic aromatic compounds): a general class of chemicals that include PAH (see below) and other organic compounds.

PAH (Polycyclic aromatic hydrocarbons): a class of very stable organic molecules. They are found in petroleum products and are formed by the incomplete combustion of wood and fuels.

Petrogenic: derived from rocks.

PMD (polyethylene membrane device): a sampling device that can be used to monitor low concentrations of organic chemicals in water.

Point source: originating from an identifiable cause or location, such as a resource extraction facility.

Priority pollutants: a list of chemicals produced by the USEPA that are either toxic, build up through food chain, or last a long time in the environment. Some of these have all these attributes.

Quality assurance (QA): a variety of tasks aimed at preserving the integrity of samples and enhancing the quality of the data. QA is achieved by having standard operating procedures for all aspects of data collection, preservation of samples, laboratory protocols, data validation, and record keeping.

Replicate samples: multiple samples taken within each combination of time, location, and any other controlled variables.

Risk assessment: the process of identifying and documenting actual and perceived risks to human health or the environment, to allow further evaluation and appropriate responses.

Synthetic crude oil (SOC): a manufactured crude oil comprised of naphtha, distillate, and gas oil-boiling range material. Can range from high quality, light sweet crude to heavy sour blends.

Tailings pond: any collection of liquid effluents or wastewater drained or separated out during the processing of oil sands or other resource extraction activities.

Terrestrial ecosystem: a system on the land, as opposed to water, formed by the interaction of a community of organisms with their physical environment.

Trace metal: heavy metals that tend to occur at very low concentrations. Examples are nickel, chromium, and lead.

Transect: a strip of ground or water along which ecological measurements, e.g. of the number of organisms, are made at regular intervals.

Trend analysis: a statistical process used to assess trends over time.

Water quality guidelines (CCME): a concentration or statement for a substance that can be compared with that substance in water. It assumes that if the concentration of the substance is higher than the guideline (an exceedence), there may be a risk to aquatic life or human health.

Water body: a stream, river, lake, or pond.

Watershed: is the area of land where all of the water that is under it or drains off of it goes into the same water body.

Variable, constituent, parameter: terms given to substances or conditions in the environment; for example arsenic, dissolved oxygen, mercury.

7.0 APPENDIX: Specific comments on statistical and related problems noted in Kelly *et al.* (2009, 2010), RAMP (2010a), Hebben (2009), and Timoney & Lee (2009)

A1 Introduction

The Committee in the course of its close examination of these documents and related ones found many problems of statistical analysis. In most cases these involve issues that are not technically complex. But they are important, as re-analyses might in some cases lead to slightly different conclusions. In other cases, as with the approaches taken with the abundance data on invertebrates, the issues are not the correctness of analyses but whether the approach taken is appropriate or useful.

We offer this appendix so that relatively simple changes needed in data analysis can be effected in the future. It also is the basis for our recommendation that relevant authorities and specialists offer some statistical workshops of a refresher nature to update data analysis skills for all scientists in the region working on oil sands issues. That might help minimize distracting future debates over statistical methodology as well as have educational value for scientists in training.

A2 Issues in Kelly et al. (2009, 2010)

A2.1 Estimation of deposition within a 50 km distance of AR6.

Potential problems in their calculations are mostly acknowledged by the authors. These estimated deposition rates must be regarded as very approximate and possibly biased.

First, as suggested by the scatter about their regression lines there is much 'noise' in their estimate. That, combined with their small sample size, would have generated wide confidence intervals, had Kelly *et al.* calculated them. These are always good reminders of not to take our point estimates too seriously.

Second, at least two systematic biases may have affected estimated deposition rates. They were calculated on the assumption that wind dispersal of pollutants from oil sands operations would have been about the same in all compass directions. Yet, as acknowledged by Kelly *et al.*, winds are predominantly from the north or the south and sampling stations for snowpack were located predominantly north and south of the oil sands operations near station AR6. That situation might have biased deposition estimates upwards. On the other hand, snowpack sampling stations were located out in the open on the iced over river and tributaries. If winter winds are strong, those locations might have been subject to wind scour. Contaminant-containing surface snow could have been blown away and redeposited in forests or low-lying areas. That would tend to lead the river snowpack data to give underestimates of actual deposition rates. Local residents and snowmobile enthusiasts should have a good sense of the likely importance of this factor.

It would be worth assessing how total deposition estimates would be altered by using only sampling stations <75 km from AR6. This approach is suggested by some indication of 'flatlining' of deposition rates at stations beyond 75 km, suggesting a 'hockey stick' model rather than one of simple exponential decline with distance from AR6.

A2.2 Disconnect between means, SEs, and ANOVAs for water sample data.

For ANOVAS, data were log transformed and the P values obtained thus refer to differences among geometric means (GMs). This is standard good practice. However apparently the only means presented in the text and figures are arithmetic means (AMs). This is indicated by use of the ± ("plus/minus") notation with standard errors and presentation in figures of only the upper error bar (implying the lower one would be of same length). The same problem is found with the means, standard errors and ANOVAs found in the online Supplementary Information. The conclusions of the ANOVAs do not apply, strictly speaking, to the AMs presented.

A2.3 Determination of detection limits (DL) and statistical treatment of <DL values are unclear and problematic.

The DLs (ng/L) given for snow in Table S1 are not easily related to what the DLs might have been for snowpack values (g/m2) in Fig. 3 or Fig. S2 (Kelly *et al.* 2009). If log-scaled y-axes were routinely used in such figures, the DL could easily be shown in them. The authors state, "To be conservative and not overestimate PAC concentrations, values below method detection limits were not increased to method detection limit." For no data set do the authors indicate how many <DL values were present. It is important to have that information. If these <DL values predominated at the distant (>75km) stations, the authors' approach might have produced regression lines with steeper (negative) slopes than if they had simply replaced <DL values with the DL. That is, the approach could have produced lower estimates of total deposition. On the other hand, the conventional argument against using estimated levels that are <DL is that they are assumed to contain much 'noise' and possibly 'bias' of one sort or another. When one is just reporting means and not doing further analyses with them, one approach is just to replace <DL values with the DL, and then report geometric means as GM = <X instead of as GM=X.

A2.4 Power analyses are redundant and misreported.

These are presented in various places (Kelly *et al.* 2009: p.4; SI, p.8). It seems clear that the effect sizes (e.g. differences between means) for which power was assessed were exactly the effect sizes as estimated from the data. Such calculations are redundant because of the direct relation between P values and power *so calculated*. The greater the P value, the less the power *so calculated*. If the authors had instead calculated power not for the observed effect sizes bur rather for some pre-specified effect size of particular interest (e.g. a difference of 20% between two means), that estimate of power may be of some interest. For example, the effect size to be demonstrated could be scaled to a toxicologically relevant concentration. RAMP IT (2010) pointed out another error in that what Kelly *et al.* (2009) were terming 'power' (1 - ß) was in fact type II error (ß).

A3 Issues in Hebben (2009)

A3.1 Length of data set needed for trend analysis is exaggerated.

It is stated that, "For a statistically defensible trend assessments on a given variable, an absolute minimum of five year's worth of continuous monthly monitoring data is required …[and] a minimum of 10 years' worth of results is generally recommended to ensure the robustness of specific trend analysis methodologies" (p. 3). Such a general claim is not supportable on statistical grounds yet it is used to justify not looking closely at trends suggested by short term

data sets. The claim apparently was intended to refer only to the specific types of trend analysis used in this report.

If there is a rough monotonic trend of, e.g., the annual mean of a variable, that trend can easily be detected with fewer than 5 years' worth of data by simple linear regression. For trend analysis use of only annual geometric means often may be the most sensitive procedure, as 'noise' due to seasonal variation is reduced. Such simple approaches of course may not work where there are missing data points or too many <DL values in the data set.

A3.2 Outlier removal is questionable and procedure is unclear.

This procedure and its rationale (p. 3) are questionable. It is not stated whether the procedure was applied to data in original units or in log units, but Hebben confirms (pers. comm. to S. Hurlbert) that it was the former. The procedure will tend to censor high values more than low. This is because this type of data (concentrations, densities, abundances) tends to have positively skewed, often approximately log-normal distributions. And indeed the figures in the report showing the "seasonality" of different elements and ions do collectively show 'outliers' greater than the median to greatly outnumber 'outliers' less than the median. This results in downward biases for arithmetic means, geometric means and medians of the outlier-censored data sets.

The operating principle used by many is that "extreme" values are retained unless there are substantive reasons for believing they are the result of error of one sort or another. The procedure outlined by Hebben (2009) would have been expected to result in elimination of about 5% of their data set, mostly the highest values. Hebben (pers. comm. to S. Hurlbert) states, however, that "very few, if any, data points were removed" from the analyses, as the identification algorithm was applied flexibly and that, moreover, all "outliers" are included in the graphs in Hebben (2009).

It is also unclear how <DL values were treated in applying this procedure; it could greatly affect the standard deviation values calculated. Hebben (pers. comm. to S. Hurlbert) states that <DL values were entered into some analyses as DL/2. This is an active area of statistical research, and few reference manuals or textbooks provide good advice on the topic.

A3.3 Finding a 'significant' step change changes test for monotonic trend.

Tests were run to see if, for a given contaminant, there was an abrupt change in its values when, in 1987, there was a change in the laboratory conducting analyses. If a significant one was found, trend analyses were done separately for pre- and post-1987 data sets. Otherwise, one trend analysis was done for the whole period. For this decision Hebben (p. 312) used an alpha of 0.10 to assess "significance." That is a stringent criterion.

Assuming a step change following a change in analytical lab is probable, especially for difficult variables like contaminants, a principal question of interest is 'how sure can we be the presumed real change is in the direction suggested by the sample means?' The odds ratio for that surety is [1 – (P/2]/[P/2]. For P=0.20, that ratio is 9:1. At least for contaminants, caution might argue for always analyzing the pre- and post-1987 data sets independently.

A3.4 Corrections for autocorrelation are inappropriate.

Measurements made on a system close to each other in time (or space) will tend to be more similar to each other than those made distant to each other in time (or space). This can be called autocorrelation but it does not, by itself, call for the adjustment of *P* values carried out by Hebben (p.5). Wherever the sampling effort is appropriately distributed over the time period (or spatial unit) of interest, "autocorrelation" is a non-problem and does not justify correction of P values.

If one measured a contaminant in a river weekly throughout two years with the aim of assessing whether there was a difference among years, and if one carried out a t-test for difference between years using only the 4 data points for each January, one could call that an error due to failure to correct for autocorrelation. In each year, the 4 January measurements likely would be more similar to each other than were the 52 measurements made over the whole year. But this would more clearly and appropriately be labeled a problem of the universe or frame from which samples were drawn (January) not corresponding to the universe implied by the question asked (year).

A3.5 Detection limit issues.

Hebben (p. 6) acknowledges by citing Helsel (2006) there are potential problems in replacing <DL values with DL/2, as he does. It is not made clear how the USGS methods for trend analysis used by Hebben treat <DL values, i.e. how "<DL" is converted to a number. Hebben (pers. comm. to S. Hurlbert) states that "Tobit regression was used to analyse censored data" and that this approach takes into account the number of DL values but does assign them numerical values.

One possibly powerful way to test for time trends that could, in most instances, eliminate arbitrary ways of dealing with <DL values would be to work only median values for years (or even pairs of years) when sufficient values per year are obtained. So long as more than 50% of values are above the DL, the median of a set of values is known even if the sample mean, either GM or AM, is not. And if in fact distributions are approximately lognormal, the median is an estimate of the GM.

A3.6 Method for determination of medians is unclear.

Though the number of censored, i.e. <DL, values is indicated for each data set presented in Tables 2-4 and appendices , it is not made explicit whether medians were determined with or without the <DL values present. Hebben (pers. comm. to S. Hurlbert) states that it was the former. Given that, a comparison of the median 1987-2007 values for key contaminants (dissolved) for the 4 stations shows marked increases in concentration (up to 200-fold) downriver that merit examination.

A3.7 Long-term trends inadequately assessed.

Given all the issues raised above, Tables 5-8 do not seem to be completely reliable summaries of information on trends. Additionally the presentation reflects the 'black-and-white', $P>0.05$ vs. $P<0.05$, approach to significance assessment. Thus a result with $P=0.06$ is treated the same as a result with $P=0.60$. The tendency for hypercondensation of information when data on so many

variables are being presented in a single document is at work here. The best sense of trends is sometimes obtained directly from the figures plotting the time series, without the aid of P values.

A3.8 Treatment and interpretation of arsenic data.

The treatment of post-1987 dissolved and total arsenic concentrations exemplifies some difficulties with the general statistical approach in the report.

In Figure 147, possible upward trends in total As are shown for both Athabasca and Old Fort. They suggest perhaps a doubling of concentrations since 2000 and a trajectory leading to guideline exceedences in the near future. Yet we are only told that "significance" is "none" because P>0.05. For all we know, P=0.056 in both cases.

For total As at Hinton and Fort McMurray and for dissolved As at all four stations (Figures 149, 150 & 151), Hebben states, "Data are insufficient for trend analysis at this time." That may be true for the particular methods used by Hebben but other approaches are available. Some of these graphs provide suggestive evidence of distinct trends, including a post-2000 decline in dissolved As at Hinton at the same time dissolved As is increasing at Old Fort.

A4 Issues in RAMP (2009, 2010)

A4.1 Monitoring and science are ill-served by the Water Quality Index.

The abstract Water Quality Index (WQI) is not scientific, is not useful for communicating to non-scientists, is distracting, and fosters superficial interpretations and summaries. The construction and weak rationale for it are given in RAMP (2009, p. 3-63). What decision-makers and the general public need from scientists is clear description of the problems possibly presented by trends in key individual measures of water quality or contamination threat, taken one by one. Simplistic interpretations of a subjectively constructed abstract index do not inform.

A4.2 Insufficient focus on patterns and their causes.

RAMP focuses almost exclusively on exceedances and whether or not contaminant concentrations fall within "historical baseline ranges." These reasonably constitute RAMPS 'alpha' level of analysis, but the more important 'beta' and 'gamma' analyses that are needed scientifically and should be a stronger focus. Spatial and temporal trends exist in all the data sets. The objective should be to characterize these as clearly as possible, and then to try to separate out the contributions of oilsands operations and other influences, such as climate change, groundwater inputs, the increasing human population of Alberta, etc.

A4.3 Inadequate approaches to assessment of benthic invertebrate communities.

These assessments are carried out with a focus on density of all invertebrates (individuals/m^2), a number of abstract indices (species richness, Simpson's index, species evenness, percent EPT), and correspondence analysis (see RAMP 2009, pp. 3-76 - 3-78). For reasons given below, these very abstract approaches are not at all useful. What is missing is serious attention to documentation of the spatial and temporal trends in the abundance of the major taxa or dominant species. About such trends we might be able to say something sensible about biological changes

in the systems. If mayflies had undergone a 50% reduction over the last ten years, there might have been no reflection of that by these abstract indices.

A4.4 Documented trends in species richness and species values are uninterpretable.

As has been known for decades, these two indices are strongly influenced by sample size (i.e. number of specimens examined or counted). These measures are not useful for any meaningful comparisons, except possibly when richness is expressed as a function of sample size by rarefaction curves.

A4.5 Suggested interpretation of high species diversity is simplistic.

It is suggested that "higher diversity and evenness are considered an indication of better conditions" (RAMP 2010, p. 3-77). There is no basis for such statements. If mayflies were the dominant species - and also the species most sensitive to a contaminant - increased levels of that contaminant, by reducing mayfly abundance, could easily *increase* the value of Simpson's diversity index. Such an increase would be nonsensical grounds for concluding that conditions were "better"!

A4.6 Correspondence analysis ordinations: a method in search of a question.

This sort of multivariate analysis is carried out both for invertebrate (e.g. RAMP 2010, p. 5-44) and fish (e.g. RAMP 2010, p. 5-65) assemblages. It leads nowhere and is a distraction. It will impress only the gullible and those afraid to question 'scientific authority'…including, unfortunately, many scientists. Its use signifies a loss of focus. Discontinue!

A4.7 Taxonomic composition by percent contribution is not sufficient.

Tabulations of the percent contribution of each major taxon to total invertebrates (e.g. RAMP 2010, p. 5-42) provides valuable insight especially when shown graphically (as for the fish: e.g. RAMP 2010, p. 5-60). However to be comprehended, the temporal patterns in absolute abundance of the individual major taxa also need to be shown.

This is especially true given that body size or mass varies by more than two orders of magnitude from one species of invertebrate in the assemblage to another.

Consideration might be given to whether estimates of mean biomass per individual might be used to provide estimates of biomass abundance for each taxon and for invertebrates collectively. Otherwise we are left with one clam being treated as the equivalent of one copepod.

A4.8 The fish data: optimizing efforts and being realistic.

The difficulties of obtaining good information on abundances of different species of fish merit close examination. Fish abundance is highly variable in space and time even in the absence of disturbing factors such as contamination from oil sands operations (e.g. RAMP 2010, pp. 5-62 - 5.63). Additional variability in data sets is introduced by the vagaries of catching or monitoring fish in the field. With this 'noise' and the current sampling effort, large reductions in the abundance of a fish species due to oil sands operations would likely be statistically undetectable

until long after the damage was done. Current efforts require significant manpower but yield small sample sizes of fish.

If additional resources are available, consideration might be given to putting them into focusing on getting better estimates of contaminant levels in fish (and other vertebrates) in the lower Athabasca River and its delta, and perhaps better data on tumors, lesions and other indicators of poor fish health as well. While interpretation of condition and contaminant levels is often complicated by the migratory or wandering habits of fish, such data are likely to be a more sensitive index of trends in fish health and threats to human health than are the inevitably 'noisy' data on fish abundances.

A5 Issues in Timoney and Lee (2009)

This is a wide-ranging review paper, but here we comment only on their analysis of the 1999-2007 data from earlier RAMP reports for PAH levels in sediments of the Athabasca River delta.

They plotted PAH values for all four delta sediment sampling stations versus time in a single scatterplot. They then calculated the Pearson product-moment correlation coefficient ($r = 0.38$) and assessed its level of significance ($P = 0.03$).

That statistical approach ignores the nested structure of the data set and has strong potential for the confounding of temporal and spatial variation. A better approach would be to do a separate linear regression analysis (PAH vs. time) for each of the four delta sampling stations. The mean of the four slopes or regression coefficients could then be tested against the null hypothesis that the true slope is zero.